MASTER'S THESIS

Towards effective Agile Data Science

de Jong, K. (Kevin)

Award date:
2019

Towards effective Agile Data Science

Degree program:	Open University of the Netherlands, Faculty of Management, Science & Technology Business Process Management & IT Master's Program
Course:	IM0602 BPMIT Graduation Assignment Preparation IM9806 Business Process Management and IT Graduation Assignment
Student:	Kevin de Jong
Date:	July 2019
Thesis supervisor	Prof.dr.ir. Remko Helms
Second reader	Jeroen Baijens, Msc
Version number:	AF200519
Status:	Final version june 2019

Abstract

Despite emerging possibilities to gain value from Knowledge Discovery, organizations are starting Data Science project that have a little chance of success, as current failure rates show us. Facing several impactful developments, such as the proliferation of Big Data, significant adjustments on traditional methodologies would make sense. Agile has the potential to inspire new artefacts to better connect Data Science activities with nowadays requirements.

The purpose of this study is to seek for contribution to the shift towards effective Data Science activities by testing the potential of an Agile-inspired design. A design science research was chosen as research method. For evaluating the proposed design, two case studies were conducted.

This research introduces the DataOps methodology, an agile inspired way of working that helps teams working within the field of Knowledge Discovery improve their results. It can be concluded that the proposed methodology has potential to move current methodologies towards, as is title of this research is called, effective Agile Data Science.

Key terms

Data Science, Agile, Methodology, Knowledge Discovery, DevOps, DataOps, Design Science Research

Summary

Data availability and ways to convert data into value is growing, just as the need for timely and effective information. Although attention for data analytics and modelling methods, called Data Science, is growing as well, the shift towards knowledge discovery from (big) data comes with opportunities and challenges. Waterfall-oriented methodologies will not provide guidance towards effective Data Science. In the field of Knowledge Discovery, KDD (1996) can be seen as the initial approach and CRISP-DM (2000) as the most commonly used Data Science methodologies. Just as data itself, possibilities to convert data into information and the need for information have been changed enormously. Despite this, survey results show Knowledge Discovery teams (still) either use methodologies founded before these shifts or they use self-invented methodologies. A proper solution is not available yet, based on the fact that nowadays 85% of big data projects fail.

In the early 2000's, a flexible, adaptive, face-to-face and knowledge sharing way of working became known, being agile. Although first orientations has been done to incorporate the advantages of agile methodologies into the KD Process, building on results from the field of Software Development, research in the field of Information Systems is mostly explanatory and not often applicable to the solution of encountered problems. Additional exploratory research is needed in order to cope with the combination of more and different data, the need for new team roles, team dynamics both within and between phases in the process, changing requirements, and a growing wish to iterate based on first results. Found literature only provide a solution on (a set of) these aspects, but fail to cover them all.

DevOps, as a concept under the "agile umbrella", show potential to deal with said developments. Many agile teams have comparable skills, rather than comparable goals, products or services that they are working on. DevOps stresses more on the communication, collaboration and knowledge sharing between developers and operators rather than tools and processes.

The DataOps methodology combines lessons learned from other research fields in order to cope with requirements for effective Knowledge Discovery. The methodology started with a presentation to a group of researchers. After some adjustments, the methodology is presented to experts, currently working within the field of Knowledge Discovery. During these interviews, a wide spectrum of evaluation criteria is reviewed. Their feedback has led to further optimization of the methodology. As a last step, interviewees were ask to provide their responses on the research findings. All interviewees confirmed their satisfaction with the methodology.

We can conclude that DataOps has added value in order to move towards effective agile data science. This research separates two groups based on their current way of working, being agile or not agile. The interviewees that are not working agile show bigger improvements than the other interviewees do, but both groups show improvements in satisfaction. This indicates that the added value not only comes from the agile aspects, but from the combination of fundamentals, process steps and model as presented as the DataOps methodology. Next to that, interviewees responded positive on a wide spectrum of evaluation criteria. This indicates that is not only covers said (set of) aspects, but has potential of being a general solution for doing effective Knowledge Discovery.

Contents

1. Introduction

1.1 Background

The availability of (diversified and unstructured) data is growing exponentially. Technical developments to store and process data (Chen & Zhang, 2014; Dragland, 2013) bring further possibilities to convert data into information and information into value, being the data value chain (Miller & Mork, 2013). Strategic data analysis and modelling methods, called Data Science, introduced a "new economy, as evidenced by large private enterprises such as Facebook, Google and Alibaba" (Cao, 2017).

Big data has great potential value. It is regarded as "the new Intel Inside, or new oil and strategic asset, and drives or even determines the future of science, technology, the economy, and possibly everything in our world today and tomorrow" (Cao, 2017). It is hypothesized that 'Big data' might lead to a form of science that is completely data driven, potentially offering a fourth scientific paradigm (Shen, 2018).

Data Science (DS) helps organizations to work smarter and make better decisions from (Big) Data (Larson & Chang, 2016; Popovič, Hackney, Coelho, & Jaklič, 2012). Several researchers presented that timely and effective knowledge availability is of great essence for organizations in their purpose to succeed and meet their business goals (Brynjolfsson, Hitt, & Kim, 2011; Larson & Chang, 2016; Pirttimäki, Lönnqvist, & Karjaluoto, 2005; Vidgen, Shaw, & Grant, 2017). Search term analysis (Google, 2018a), growing shortage of specialists (Bowley, 2017; Davenport, 2012; Piatetsky, 2018) and growing spend on big data analytical solutions (Goepfert & Shirer, 2018) confirms the continuous growing attention from businesses towards knowledge discovery. KD methodologies guide the process of gaining knowledge out of data (Alnoukari, Alzoabi, & Hanna, 2008; Larson & Chang, 2016; Li, Thomas, & Osei-Bryson, 2016; Lim et al., 2018).

The shift towards knowledge discovery from (big) data comes with both opportunities and challenges (Zhou, Pan, Wang, & Vasilakos, 2017). The success of a KD project is highly depending on (the cooperation between) people, processes and technology (Gao, Koronios, & Selle, 2015). The changing characteristics of data (Larson & Chang, 2016) have created separated, specialized roles within different phases (J. Saltz, Hotz, Wild, & Stirling, 2018; Thomopoulos, 2018). Results and human assumptions influence following phases (Sumana Sharma & Osei-Bryson, 2009) potentially leading to a product that does not match end-user expectations (Larson & Chang, 2016). An extra challenge comes from changing requirements during the process, originated from rapidly evolving environments (Givanildo Santana do Nascimento & de Oliveira, 2012; Wilkes, 2012). Having clear requirements is difficult in the first place, due to their "exploratory and often ad-hoc nature" (Das, Cui, Campbell, Agrawal, & Ramnath, 2015). Traditional waterfall-oriented methodologies are not an adequate answer to these challenges (Das et al., 2015; Gil & Song, 2016; J. Saltz, 2015). However, they are still frequently used (KDNuggets, 2014).

Currently, 85% of the big data projects fail (Walker, 2017), with 60% fails to even go beyond the pilot and experimentation phase (Gartner, 2015). Within Software development (SD), changing requirements, high costs, long lead-times and high failure rates resulted in the development of agile working methods (G.S. do Nascimento & de Oliveira, 2012). The Manifesto for Agile Software Development (Beck et al., 2001) has fundamentals that can potentially deal better with increasing nature, scale and dynamics of knowledge discovery from big data (Vidgen et al., 2017).

During the years of impressive developments and practical solutions in the big data environment, current KD methodologies have not evolved significantly from the step-by-step KDD process, which were already founded in 1993 (Mariscal, Marbán, & Fernández, 2010; J. S. Saltz, Shamshurin, & Crowston, 2017). Survey results, referred by Schmidt (2018), Alnoukari (2012) and others, show a wide use of CRISP-DM (43%), SEMMA (17% and the KDD Process (7,5%) methodologies (KDNuggets, 2014) for knowledge discovery. Noteworthy is the increase in people using their own methodology, rather than using the KDD methodologies (J. S. Saltz et al., 2017). This can be seen as an indicator that existing methodologies don't suit contemporary requirements, resulting in high failure rates. Objective of this research is to further understand how elements of agility can contribute to improvement of the knowledge discovery process (KDP).

1.2 Problem statement

An exceedingly high rate of KD projects fail. Several developments have drifted traditional methodologies away from current requirements. There is need for a methodology that can adapt these developments in order to achieve an effective knowledge discovery process.

1.3 Research objective and questions

One can recognize various process methodologies used for data mining and knowledge discovery (Alnoukari & El Sheikh, 2012). Not having a central methodology, nor a handful of methodologies with defined use-areas, indicates the search for a methodology capable of working with contemporary requirements in order to solve said high failure rate.

This research will provide an overview of the main existing KD methodologies. Since their value shifts over time (Shen, 2018), this research will set emerging trends in the field of KD against the methodology characteristics and will search for valuable lessons from other research fields. For example, within SD, The Manifesto for Agile Software Development (Beck et al., 2001), shifted methodologies from a waterfall oriented process towards a more flexible and adaptive process. The knowledge discovery process (KDP) can potentially leverage from both the fundamentals of agile and lessons learned within SD, and hence, bring failure rates down.

In order to get an overview of the value of agile principles in KD Processes, this research seeks for answers to the following main- and sub questions:

How can agile methodologies contribute to the knowledge discovery process within organizations?

To get a clear answer to the research question, the sub-questions to answer are:

1 Which knowledge discovery methodologies are used within the knowledge discovery process, what are their characteristics and to what extent do these methodologies suit current requirements?

2 What are the different characteristics of agility?

3 Which agile knowledge discovery methodologies have embodied agile characteristics?

4. In order to determine the value of the purposed KD methodology, what can be considered as strengths of the models and what aspects leave room for further improvements?

1.4 Motivation/relevance

Over the last decade, many scholars have highlighted the need for empirical investigations on KD (Côrte-Real, Oliveira, & Ruivo, 2017; Dingsøyr, Nerur, Balijepally, & Moe, 2012; Zeng & Glaister, 2018). A gap exists between the growing attention to, and potential of, KD and the lack of a proper methodology guiding towards effective KD (J. S. Saltz et al., 2017). Traditional methodologies will not fill this gap (Das et al., 2015; Gil & Song, 2016; J. Saltz, 2015). Based on lessons learned from other fields, first scientific orientations have been done on the advantages to incorporate agile in the KD Process (Jeffrey Saltz, Heckman, & Shamshurin, 2017). Since research in Information Systems is mostly explanatory and not often applicable to the solution of encountered problems (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2008), this research seeks for a proper reaction on said problem statement. Building further on existing knowledge and current requirements, this research contributes to science by testing an agile-inspired methodology within the field of data science.

As a result, this research will provide observations that can be useful for businesses to improve their KD activities and extends base for further research.

1.5 Main lines of approach

After the formulation of a problem statement and research questions, the next step towards establishing a successful research was building a theoretical framework on which further work can be constructed. This can be found in chapter 2. Actions are undertaken to build said agile-inspired methodology and test it empirically.

Based on the four main purposes for scientific research, being exploratory, descriptive, explanatory and evaluative purposes (Saunders, Lewis, & Thornhill, 2016) this study has mainly an exploratory purpose. Chapter 3 describes the empirical part in detail, which is guided by the Design Science Research approach (DSR). DSR focuses on the development and performance of methodologies.

2. Literature review

The next step is to do a systematic literature review. A pre-planned strategy will be used to find existing literature to analyse how this literature provides answers to the research questions and, in a wider perspective, how they can already help filling the gap of a methodology suiting contemporary requirements, as defined above. During this research, the *eight-step guide* of Okoli (2010) is used since it is built to meet the unique needs of information systems (IS) research. The following steps have been taken:

1. Purpose of the literature review (2.1);
2. Protocol and training (2.2);
3. Searching the literature (2.3);
4. Practical screening (2.4);
5. Quality appraisal (2.5);
6. Data extraction (2.5);
7. Analysis of findings (2.6) and
8. Writing the review (2.6).

Research approach

2.1 Purpose of the Literature Review

The literature review forms a “systematic, explicit, comprehensive and reproducible method for identifying, evaluating and synthesizing the existing body of completed and recorded work produced

by researchers, scholars and practitioners" (Okoli & Schabram, 2010). It helps in formulating conclusions about the knowledge available from previous work, brought together based on their contribution on the sub-questions 1a to 2b of this research. The literature review indicates, as a main purpose, direction to the empirical research and brings provisional answers to the research questions. Based on that foundation, an agile-inspired appliance is outlined to test empirically, following the methodology of Peffers et al. (2007).

2.2 Protocol and training

Considering protocol and training, step 2 of Okoli's structure in literature reviews (2010), there is no relevance to implement any team-techniques since the research is done by a single researcher.

Implementation

2.3 Search for the Literature

The research question **"How can agile methodologies contribute to the knowledge discovery process within organizations?"** is used to select main nouns, which are connected using Boolean logic. Where possible, synonyms where added to the query.

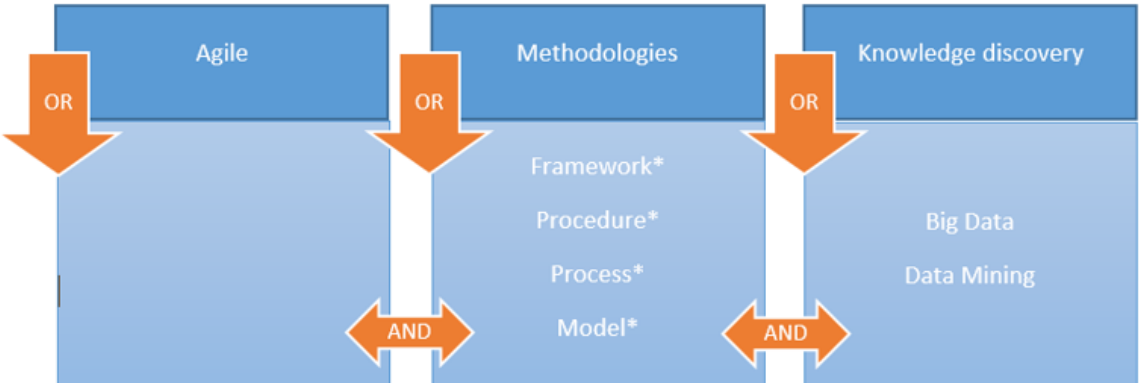


Table 1: definition of the search string used for the literature review

By putting the nouns between quotation marks, articles that had subject words (e.g. knowledge and discovery) somewhere in the article were excluded, resulting in more relevant articles. A further match with the research goals was met by setting the rule of "methodologies", "knowledge discovery" and synonyms being part of the abstract. The choice is made not to do this for agile, since the results reduced too far and agile as a concept can be introduced later in the articles, i.e. after introduction in the abstract.The final query is:

> ((Abstract: ("**Knowledge Discovery**")) OR (Abstract:("Big Data")) OR (Abstract:("data mining")) OR (Abstract:("data science")) OR (Abstract: ("KDD")))
>
> AND ((Abstract:(**Methodolog***)) OR (Abstract:(Framework*)) OR (Abstract:(Procedure*)) OR (Abstract:(Process*)) OR (Abstract:(Model*)))
>
> AND (**Agile**)

With the search string defined, the search criteria help to guarantee a comprehensive, explicit and reproducible research (Okoli & Schabram, 2010). The search criteria are as follows:

a. The Open University's Library is selected to find peer-reviewed articles;
b. The disciplines of focus are "Business" and "Computer Science";
c. The article is relevant and useful for the research, based on the abstract;
d. The search terms used are obtained from the research questions;

e. Only articles published after 01-jan-2010 to 01-nov-2018 will be in the scope;

Aspects of the protocol are selected based on (a) quality and availability, (b) relevance of the discipline, (c) contribution to the research (d) link with the research question, (e) actuality.

The robustness of the research will grow by both backward and forward snowballing via Google Scholar. It helps to find additional papers as well as it provides a view of articles building on the key articles found in previous steps (Watson & Webster, 2002).

Search on:	*Number of results:*
Nouns (not using quotation marks)	6584
Nouns between quotation marks	1009
Nouns are defined in abstract	133

Table 2: search results by different approaches

2.4 Practical screen

For the practical screen, only articles that match the discipline, contribute to the research and that have a link with the research question. Excluding these articles resulted in a focus research, which can build on the information needed in order to lay a good foundation.

A total of 133 articles were found by using the query, 128 where available and 118 seemed relevant based on their title. For 61 articles, the article was considered relevant to the subject, based on the abstract.

Criteria:	*Articles: (133 = 100%)*
Available via Library Open University	128 (96%)
Relevant title, downloaded to read the abstract	118 (89%)
Relevant abstract, selected for quality appraisal	61 (46%)

Table 3: practical screen, filtering towards useful set of articles

Next to the articles found by using the search query (Attachment 1a), eleven articles (Attachment 1b) are provided by the University, bringing the total to 71 selected and unique articles. Within said articles, 10 references, not found via the query, are considered contributing to this research. They can be found in Attachment 1c.

2.5 Quality appraisal and data extraction

All articles are screened for their claims, the evidence supporting them, if the evidence is warranted and how it is backed. This can be theory or a case study. Reading the research approach was done, including the limitations of the research. Articles matching the quality criteria are considered relevant and useful for this research. All articles are considered of good quality.

Information is extracted from the scientific articles left from previous steps. They are reviewed for touch points with main subjects of this research, like agile, KD methodologies and the knowledge discovery process. Found information, by reading the full article, is clustered. This is done based on their connection to the defined questions. Articles can contribute to more than one research question. The combinations can be found in Attachment 1a, 1b and 1c.

Several articles seemed relevant by reading the abstract. Reading the full body of the research confirmed differently, matching the research subjects just partly. A total of 15 articles is used for further research steps.

Results and conclusions

2.6 Synthesis of studies and writing the review

Existing methodologies for Data Science

In the field of data science, a wide variety of process models for the approach and execution of analysis projects emerged (Alnoukari & El Sheikh, 2012). KDD can be seen as the initial approach and CRISP-DM as a central approach of the evolution towards the most commonly used KD methods (Mariscal, Marbán, & Fernández, 2010). KDD involves the broadly seen, sequential steps of data selection, pre-processing, subsampling, transformation, data analysis post-processing and knowledge utilization (Shen, 2018). CRISP-DM introduced new steps: business understanding and data understanding, "two cornerstones of any successful data mining project" (Alnoukari & El Sheikh, 2012) and deployment (Mariscal et al., 2010). The high level guidelines and lack of attention for stakeholder dynamics are considered as weaknesses of the CRISP-DM methodology (Mariscal et al., 2010; Vishakha Sharma, Stranieri, Ugon, Vamplew, & Martin, 2017). Further improvement can be reached by implementing maintenance activities dealing with new data observations and software- and model updates (Mariscal et al., 2010). The initiative to renew the CRISP-DM towards CRISP-DM 2.0 in order to align with the first aspects of Big Data, as mentioned by Mariscal (2010), entered a frozen state since early 2007 (Li, Thomas, & Osei-Bryson, 2017), confirmed by a search in the University library.

Including CRISP-DM, Attachment 2 shows an overview and comparison of the most popular knowledge discovery methodologies and their steps. In order to get an overview and comparison of similarities between methodologies, comparable steps are positioned next to each other. The presented methodologies are KDD, SEMMA, Six Sigma, Cabena et al., Anand & Buchner, Knowledge Discovery Life Cycle, CRISP-DM, Cios et al., the Snail Shell Process Model and Combined dual-cycle methodology (Alnoukari et al., 2008; Azevedo & Santos, 2008; Hofmann & Tierney, 2009; KDNuggets, 2014; Kurgan & Musilek, 2006; Li et al., 2016; Mariscal et al., 2010; J. Saltz & Heckman, 2018; Shen, 2018).

Knowledge Discovery Methodologies and their fit in the era of Big Data

A fairly static, pure and small dataset and a certain searching direction, being small data, can be analysed by a single person (Shen, 2018). A methodology is not critical here. Since the amount of diversified data(sets) expands fast, entering the era of Big Data, a team of specialists replaced this single person to do the job. Such teams need a methodology for effective KD (J. Saltz, 2015). Big Data, characteristics of data are not only velocity, variety and volume, but also value, variability and veracity (Janssen, van der Voort, & Wahyudi, 2017; Sugam Sharma, 2016). Although Big Data attracted much attention from business and science, generally accepted definitions and key concepts of Big Data are lacking (Mikalef, Pappas, Krogstie, & Giannakos, 2018). Within this research, Big Data is defined as lots of (near) real-time, diversified data (Li et al., 2016; Oztemel & Gursev, 2018), useful for valuable and actionable Knowledge Discovery (Cao, 2015; Shen, 2018).

Not much research has been conducted on the fundamental differences between processing small and big data (Shen, 2018). This could be the reason why methodologies have not evolved significantly from the step-by-step KDD process (Mariscal et al., 2010; J. S. Saltz et al., 2017). Traditional methodologies are not adequate for solving challenges arising from Big Data (Gil & Song, 2016). Chen & Zhang (2014), Kaisler, Armour, Espinosa & Money (2013) and Labrinidis & Jagadish (2012) define challenges within individual methodology process steps. Next to these, major challenge is the high dependencies between the numerous phases and tasks. Results and (human) decisions from previous phases influence, obviously, following phases (Matsudaira, 2015; Sumana Sharma & Osei-Bryson, 2009). Formulating a rock-solid objective by the end-user, as a first step of the KD-process, is already a major issue. The (re)formulation of a problem statement depends on e.g. the available data, resources, knowledge and changing (environmental) circumstances (Di Orio,

Cândido, & Barata, 2015; V. Sharma, Stranieri, Ugon, Vamplew, & Martin, 2017). Parameters change during the process based on new data and/or obtained insights. Since previous phases influence following phases, checklist based waterfall methodologies with only results at the very end of the process don't suit well (Davenport, 2014; Grady, Payne, & Parker, 2018; Hofmann & Tierney, 2009; Sumana Sharma & Osei-Bryson, 2009; Wilkes, 2012).

First steps from traditional methodologies towards a more iterative process are taken by connecting the finishing step with the first step, forming a loop. This provided the opportunity to redo the process with new, valuable parameters. With the need for maintenance of the KD-process (Li et al., 2016; Mariscal et al., 2010), this research takes the position of KD rather as continuous delivery than as a project, having a defined end-phase. CRISP-DM (2000), methodologies like Cios (Cios & Kurgan, 2003) and the Snail Shell process (Li et al., 2016) separate from other methodologies having several significant loops in the process. This not only interconnects steps, but also experts. Although this sounds promising, Hofmann & Tierney (2009) show that KD-experts are facing unstructured and inefficient communication and documentation. The survey of Ho (2017) provides five main challenges for data professionals, all subject to process steps where these specific professionals, according to Hofmann & Tierney, have no role in.

Although 82% of data scientists don't follow a process, 85% of experts expect better results with activities guided by a suiting methodology (J. Saltz et al., 2018). Despite this need, looking at all 296 papers from the IEEE Big Data Conference 2014, ironically, none was focused on methodologies and tools for improved effectiveness (J. Saltz, 2015).

Agile characteristics and agile-inspired methodologies

Agile methodologies appeared in SD, rather having intense end-user involvement, fast delivery and end-user satisfaction than rigid systems and checklists (Ben Ayed, Ltifi, Kolski, & Alimi, 2010). The solitary task of developing software, focusing on delivering projects in some steps made room for attention for group dynamics and an iterative process, being main concepts of agile (Balle, Oliveira, Curado, & Nodari, 2018; J. S. Saltz et al., 2017). Agile as a widely accepted working method for SD provided improvements in delivering on time, within budget and meeting customer expectations (Brhel, Meth, Maedche, & Werder, 2015). By following the ideas of the Agile Manifesto (Beck et al., 2001), SD became less formal, more dynamic and customer focused (Larson & Chang, 2016), all useful and promising aspects to reduce failure rates within DS. The dynamic and adaptive Speculate, Collaborate, Learn lifecycle replaced the static steps of Plan, Design, Build (Alnoukari et al., 2008; Highsmith, 2000). Continuous delivering work encourages feedback, what can be incorporated into any future decisions (Grady et al., 2018). One can say the concise Agile Manifesto, containing (just) four values and 12 principles, has introduced a new paradigm. And yet being concise, literature lacks clarity as to what defines agility. It seems almost every piece of research adopts a unique interpretation of agility (Abrahamsson, Conboy, & Wang, 2009). We have seen the introduction of several methods that appeared from that fundament including Scrum, XP, Kanban, Lean SD, feature-driven development (FDD) and Crystal (Dingsøyr et al., 2012). Based on found literature, Scrum, XP and Kanban are regarded as most frequently called agile-inspired methodologies. They are compared in detail in Attachment 3. For Scrum and Kanban, evidence is found for adaptation of these agile methodologies in KD, as described in the next section.

For DevOps, another concept found under the "agile umbrella", characteristics showed promising capabilities on noticed developments in KD. DevOps' footprint within currently available DS research is found small. The concept of DevOps is relatively new, launched in 2008 (Hemon, Monnier-Senicourt, & Rowe, 2018) and peaking during this research (Google, 2018b), but having "a paucity of DevOps research directly related to IS" (Sharp & Babb, 2018). Generally, based on publication

subjects, agile methods have passed early innovation adoption phase and become mainstream, where DevOps obtains more attention (Mishra, Garbajosa, Wang, & Bosch, 2017). Many organizations make agile focus teams, according to Conway's Law mistakenly splitting them functionally based on technology, rather than on product or service (Debois, 2011). DevOps extends agile in terms of the principles, since DevOps stresses more on the communication, collaboration and knowledge sharing between developers and operators rather than tools and processes. DevOps can achieve agile goals to reduce team working latency (Hemon et al., 2018; Jabbari & Ali, 2016). Humble and Molesky (2011) highlighted the core values of DevOps, being culture, automation, measurement and sharing. Within this study DevOps is defined, based on Jabbari & Ali (2016), as a methodology aimed at bridging the gap between Development (Dev) and Operations (Ops). It emphasizes communication and collaboration, continuous integration, quality assurance and delivery with automated deployment utilizing a set of practices. DevOps reduces mutual tension of creativity and speed (build) versus quality, stability and traceability (run) (Hemon et al., 2018; Jabbari & Ali, 2016). DevOps lacks a manifesto and a consistent prescriptive methodology and is still evolving (Babb, Nørbjerg, Yates, & Waguespack, 2017). Although the case study of Hemon et al. (2018) shows both promising results and attention points, more research is required to uncover DevOps' full value in Data Science.

Data Science embodying Agile

Vidgen, Shaw and Grant (2017) have, based on a variety of 60 DS professionals, identified numerous challenges organizations face in creating value from big data and analytics. Also, they state that DS can learn much from agile SD like engagement with the end-user, frequent delivery, iterations, colocation of specialists and rotation of roles. Analytics should iterate in order to grow in fidelity, stopping when the results are sufficient to meet outcome requirements (Grady et al., 2018). The Agile Manifesto "values learning and self-empowered teams that reflect upon and improve their skills and practices on an ongoing basis" (Babb et al., 2017). Two main DS agile methodologies that embodied this principle are Agile Scrum and Agile Kanban, created in respectively SD and Lean Manufacturing (Lei, Ganjeizadeh, Jayachandran, & Ozcan, 2017; J. Saltz & Heckman, 2018). Scrum's concepts are the user story, sprint backlog being the completed work during the sprint, product backlog being a wish list for next sprints, the one or two week lasting mini-project sprints and the update meetings, known as daily scrums (Larson & Chang, 2016). Kanban is "simple in structure", build for limiting the amount of work in progress to activities primarily focused on delivering customer value (Brechner, 2015). Instead of completing lots of work in phase A and hand it over, the amount of work in phase A is aligned with the following phase(s) to get efficient delivery (Lei et al., 2017). Scrum and Kanban are highly adaptive, but Scrum is the most prescriptive of the both. Saltz & Heckman (2018) did a quantitative experiment comparing Crisp-DM, Scrum and Kanban, unfortunately not showing significant improvements delivered from both agile methodologies. New models, such as proposed by Li et al and Shen, present new, agile-inspired models, but do not zoom in to relations and communications between phases and professionals.

2.7 Summary and objective of the follow-up research

Above shows the changing requirements for dealing with data due to changing data characteristics as a result of Big Data and stakeholders' dynamics. Empirical research on how to do this in the field of DS lags behind. Since phased waterfall-oriented processes are making room for iterations in the process, communication and mutual knowledge sharing becomes key. The heavy-lifting done after implementing the Agile Manifesto has laid the foundation for subsequent work. Although said improvements delivered from agile, popular agile methodologies as Scrum and Kanban have not broken down silos between build and run. This is based on the fact that its focus is on the methods

and practices of the build-area (Hemon et al., 2018). DevOps takes a higher level and larger scope and can be seen as an extension of agility.

Objective of the follow-up research is to present and empirically test a model that extends existing methodologies and incorporates Agile DevOps characteristics that can be of value towards effective DK.

3 Methodology

After completing previous literature review, requirements towards successful knowledge discovery are formed. The empirical part of this research intents to build further on said scientific, literature foundation, further contributing by adding an empirically tested model.

3.1 Research methods

The literature review provided valuable reasons for building an artefact and requirements of how to build such an artefact. The evaluation assesses the value of an artefact (Venable, Pries-heje, & Baskerville, 2014). This research can form a step towards extended research, giving it aspects of a formative evaluation.

Research purpose and approach

Based on the four main purposes for scientific research, being exploratory, descriptive, explanatory and evaluative purposes (Saunders, Lewis, & Thornhill, 2016), this study has mainly an exploratory purpose. Obviously, a research study may fulfil more purposes. The research includes, based on the exploratory nature, a search of the literature as done earlier and in-depth semi-structured interviews. Evaluation is done to review how well the KDP artefact matches environmental expectations (relevance) and gains scientific knowledge (rigor) (Venable et al., 2014).

Instead of a cause-effect link between variables, this research explores a phenomenon by testing the potential of a design. Theory follows data, rather than vice versa, which points the research, together with the ambition to generate theory rather than verify it, towards an inductive approach (Saunders et al., 2016). Final deliverable of the research is a valuable methodology. In order to evaluate this methodology, this research intent to show the value of the methodology itself and agile in general.

Design Science Research Methodology (DSRM) as paradigm for this research

Within the field of Information Systems (IS), research output is mostly explanatory and not often applicable to the solution of encountered problems. Design science research, being “The act of creating an explicitly applicable solution to a problem” is broadly accepted in other fields. For IS, DSR has just a small share of publications, producing artefacts applicable to research or practice (Peffers et al., 2007). It separates from other paradigms (positivist, interpretivist, critical), having attention for designing, developing and building new artefacts, next to evaluation demonstrating the utility, quality and efficacy of a design artefact (Sangupamba Mwilu, Comyn-Wattiau, & Prat, 2016; Venable et al., 2014). Artefacts are the innovations helping to analyse, design, implement and use IS more effectively and efficiently via ideas, practices, technical capabilities and products (Hevner, March, Park, & Ram, 2004). DSR seeks to extend “the boundaries of human and organizational capabilities by creating new and innovative artefacts” (Hevner et al., 2004). DSR is essentially a search process to discover an effective solution to a problem (Hevner et al., 2004), in this case the high failure rates within KD.

Empirical steps

The demonstration, which contains all empirical steps teaken, started with a with a presentation to a

group of researchers. After some adjustments, the methodology is presented to experts, currently working within the field of Knowledge Discovery. As a last step, interviewees were asked to provide feedback on the result of the processing of all interviews.

3.2 Technical design: elaboration of the method

The DSRM (Peffers et al., 2008) consists of a six step process model. The groundwork for this research is laid out in the first chapter of this document. It defines the problem (step 1) and motivation for doing further research. The objectives (step 2) are creating and testing an agile-inspired methodology within the field of data science in order to deliver the described contribution. Step 1 and 2 are discussed in previous sections.

1. Problem identification and motivation;	4. Demonstration;
2. Define the objectives for a solution;	5. Evaluation;
3. Design and development;	6. Communication.

Table 4: DSR process steps (Peffers et al., 2007)

3.2.3 Design

The proposed model of DevOps stresses more on the communication, collaboration and knowledge sharing between developers and operators rather than tools and processes, as described in previous sections. Based on found literature, this research defines the principles for DataOps being continuous delivery, sharing knowledge, automation, shared responsibility, measurement and comparability. Apart from an e-book published by DataKitchen (Bergh, Benghiat, & Strod, 2019), again - having no scientific literature foundation or empirical tests, a formal Manifesto is available (Jabbari & Ali, 2016). This research provides a model of DevOps for Knowledge Discovery, the DataOps methodology.

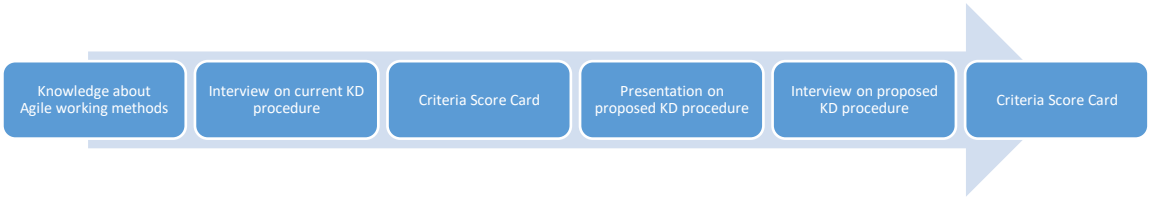
3.2.4 Demonstration

During the demonstration, the use of the artifact to solve one or more instances of the problem are tested (Peffers et al., 2007). A case study provides rich, empirical observations that lead, ultimately, to a theory (Saunders et al., 2016). Interviews were conducted individually. This setting, rather than having e.g. focus group interviews, gives the opportunity to zoom in to their vision and lets interviewees be as open as possible about that communication.

During the sessions, a total of 6 interviewees participated. Before the interviews, questions were formulated in order to guarantee uniformity. The questions are based on the intention to get to learn about the agile experience of the interviewees, their vision on their current way of doing KD and the DataOps methodology. For that vision, a general observation was asked, just as their vision on defined evaluation criteria. Participants are representatives of their teams, having different roles in the KD process. The group composition is based on an evenly distributed representation of roles.

The expert interviews contains two or more rounds, using the Delphi method. This provides interviewees the opportunity to adjust answers based on answers provided by other interviewees. The Delphi method will be explained in detail in section 4.4 (testing the design).

Within the interviews, several steps were involved. After a neutral introduction and a conversation about the interviewee's agile experience, the current way of KD will be discussed. After that, the interviewees have individually filled in a 'criteria score card' (see section 3.2.5). Next, the alternative design is demonstrated by showing the design and a neutral explanation. The explanation was put on paper to guarantee objectivity. Next, another round of questions helped to gather information. Ultimately, a new 'criteria score card' is filled in by the interviewees.



Setting of the interviews

Before participating, the interviewees received a statement about what, where and why data is stored. A participant information sheet is provided, including an anonymity- and confidentiality statement. The interviews are held in a convenient room, where it's unlikely to be disturbed, and the interviews were recorded including contextual data. Different types of questions, except leading questions, are asked and answers where summarized to test understanding. All interviewees where free to participate and withdraw at any moment.

During the research, participant validation (Saunders et al., 2016) is warranted by summarizing answers provided. After the interviews, interviewees received a summary of the interview with a request for confirmation. After the study, interviewees received research findings and observations, giving them the opportunity to comment.

3.2.5 Evaluation

In the evaluation phase, the model is tested on several requirements. These were found in the available literature and conducted from the first round of interviews. Results will be measured via both qualitative interviews and verified via quantitative evaluation. The interviews are summarized and summaries are used to extract observations on specific criteria as presented below. A questionnaire is used to do a quantitative evaluation. Ultimately, after the empirical research, a strong statement can be made about the utility (Venable et al., 2014).

1. Usable	8. Frequent delivery
2. Efficient	9. Guiding teamwork
3. Reliable	10. Iterations appreciated
4. Maintainable	11. Fast delivery
5. Flexible	12. Respecting budget (hours/money)
6. Reusable	13. Self-empowered teams
7. End-user involvement	14. Stimulates a learning-curve

Table 5: Evaluation criteria based on Mathiassen et al. (2000)

3.2.6 Communication

Based on its rigor and relevance of the research, this research provides contribution to both science and business. Being inductive, the research leaves room for further exploration on its observations. To enable sharing knowledge gained from this research, case names and interviewees are made anonymous.

3.3 Rigor and relevance

Scientific evaluation of a design concerns the artefact in the context of both its contribution to the environmental-based relevance cycle and the knowledge-based rigor cycle (Venable et al., 2014). The key purpose of evaluation is to determine how well a design achieves its expected environmental utility. Other important purposes are to determine the quality of knowledge outcomes and enable a comparison with other artefacts (Venable et al., 2014). In general, evaluation identifies "potential weaknesses in the theory or artefact and the need to refine and reassess" (Hevner et al., 2004). Rigor and relevance are not two separate concepts but basic requirements "to live DSR up to its label as science" (Venable et al., 2014).

The Framework for Evaluation in Design Science (FEDS) from Venable et al. (2014) guides DSR. The framework has 2 axes: artificial/naturalistic and formative/summative. DSR evolves from artificial/formative towards naturalistic/summative. (Venable et al., 2014). Shifting towards naturalistic evaluation, rigor increases. Another factor influencing rigor is the correlation between the artefact and the observations. This research strives to exclude external ground for observations by asking in a uniform and neutral way. The application in the appropriate environment, next to the additions to the knowledge base, both part of the dual-cycle ISRF of Hevner et al. (2004) is outlined in chapter 1. In short, the objective of this DSR is to develop a solution to an important and relevant business problem, is the high failure rate of KD activities.

4 Design

This section provides further details of the proposed Knowledge Discovery methodology. The methodology contains both key elements and design steps. It contains of 14 key principles that lead towards successful operationalization of the methodology. Implementing best practices from several agile-inspired methodologies, these principles lead to a more flexible and adaptable way of conducting knowledge discovery. The DataOps methodology is build up from inspiration from other methodologies, as presented in Attachment 5, and from insights of the researcher.

Key definitions

The methodology is built on several key definitions. In order to provide both interviewees and readers of this document with a clear understanding, several definitions are defined, being:

- Methodology:
  The methodology includes the total package of documentation (fundamentals including roles, process steps, model) that is provided within the demonstration. This package needs to provide knowledge discoverers with enough information to work with the methodology of DataOps.
- Model:
  The overview or blue print of the process steps, as provided below.
- Iteration:
  Conclusions in one process step potentially create the need to go back to a previous step. An iteration allow teams to do this.
- Loop:
  After doing all the process steps, sometimes twice when an iteration was needed, the process is completed for the first time. When finishing a loop, first results are available. If these results are not satisfactory, teams can start a new loop.

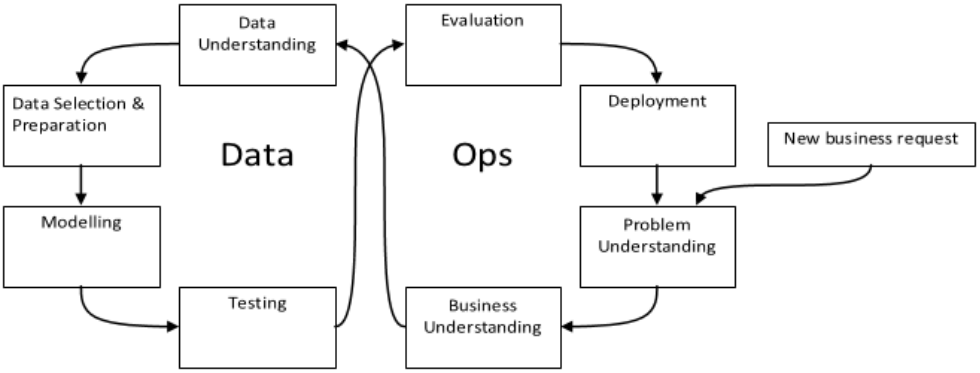
4.1 Blueprint of DataOps

General description

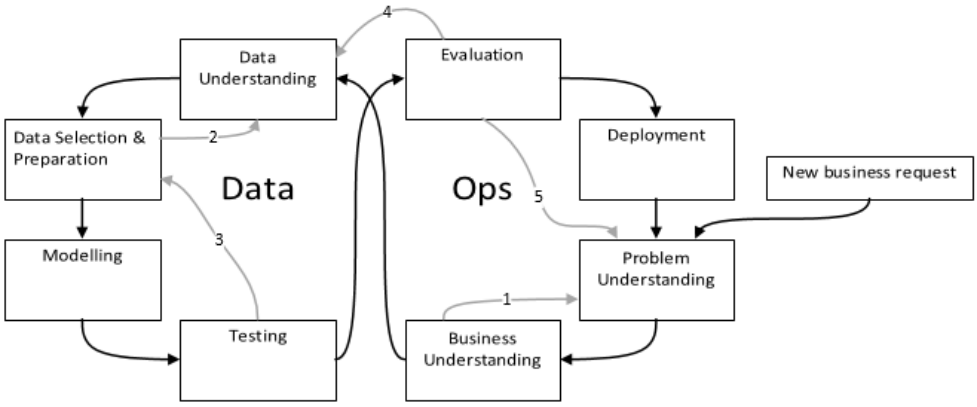
Like all Agile methodologies, DataOps is built directly on the Agile manifesto (Beck et al., 2001). Several existing methodologies are inspired by the Agile principles (Cao, 2017; Dingsøyr et al., 2012; G.S. do Nascimento & de Oliveira, 2012; Larson & Chang, 2016; Li et al., 2016), which creates the possibility to combine best practices, key elements and process steps, next to filling the gap as subject of this research. In creating the DataOps model, CRISP-DM is used as a starting point. As described earlier, the modification of CRISP-DM, based on a need to handle big data characteristics, integration with the business and closed-loop feedback and team elements (Mariscal et al., 2010), has not been updated since early 2007 (Li et al., 2017). This model aims to fill this gap by incorporating design steps of mainly DevOps, combined with other methodologies as KDD, Scrum, the Snail Shell and the Dual-cycle methodology.

The DataOps model can be explained in three layers as presented below. Model A is a basic design, showing the most lean-and-mean route. Model B creates the possibility to iterate, which enriches the DataOps design by creating the possibility to redo a specific process step without the need to redo all process steps, as needed in a rather linear process. Model C, presented as the DataOps design and presented on the next page, has that possibility to iterate and is set for welcoming new data observations into the process creating the opportunity to continuously take steps towards quality improvements of the delivered product.

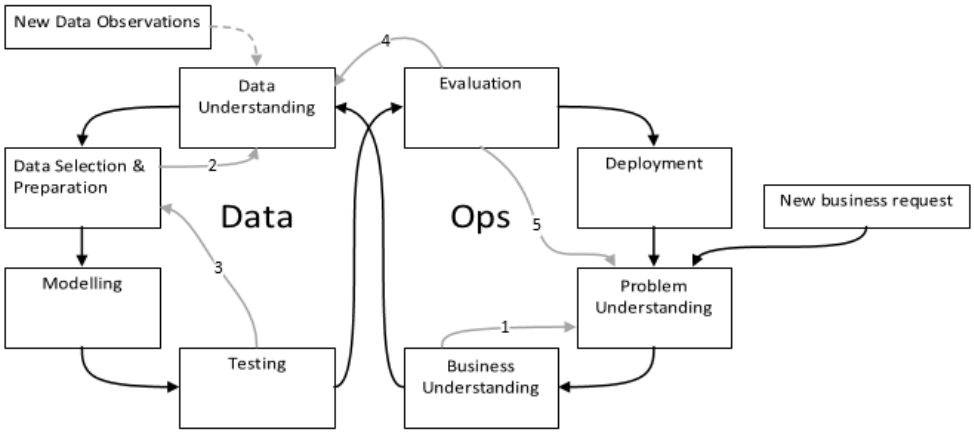
Model A: A basic version of the DataOps model



Model B: The DataOps model including iterations



Model C: The DataOps model including iterations and external influences



The methodology contains of two (Data+Ops) times four process steps. Every step takes a calendar week, meaning that without the need to iterate, a loop can be done within 8 weeks. Fast availability of first results is ensured, optimization of these results can be done by doing (a) new loop(s).

Some steps need less than a week. Teams have temporary update moments with the possibility to continue working on the next step. When more time is needed, teams need to continue. First deliverables can help the team to learn and understand each other. When starting a new loop or iteration, the team can continue working on the more time-consuming task.

Key Principles
The methodology of DataOps has several aspects of DevOps, mainly influencing the key principles. These key principles are the building blocks that provide the possibility to use the methodology successfully. As described in detail in the literature review, DevOps stresses more on the communication, collaboration and knowledge sharing with the ambition to bridge the gap between Development (Dev) and Operations (Ops). The key principles are provided in detail in section 4.2.

Design steps
Inspired by CRISP-DM, the design contains of 8 process steps, visually separated in two parts. Process steps or activities on the left side are focused on data, where process steps or activities on the right are focused on operations. Opportunities to share knowledge between both sides will mainly be found in the activities that are visually next to each other. Process- and design steps will be explained in detail in section 4.3.

Improvement on existing Knowledge Discovery methodologies
In comparison with methodologies that are not agile inspired, the methodology adds opportunities to deliver on time, within budget and in line with customer expectations by making the process more flexible and adaptive. Agile fundamentals deal better with increasing nature, scale and dynamics of knowledge discovery from big data (Vidgen et al., 2017). When comparing existing agile methodologies, the DataOps has more focus on communication, collaboration and knowledge sharing between Data- and Ops oriented team members. As well, it has the necessary focus on automation of process steps in order to keep the process efficient.

4.2 Key principles of DataOps

DataOps' key principles consist of both fundamentals and roles, all applicable during the whole process. Just as the Agile Manifesto, this methodology uses fundamentals as guidance for the teams. The process steps help to explain what to do within a specific step. They are based on the researched methodologies within the literature research and the comparison of the methodologies as presented in Attachment 2a/b and 3. Where needed, the researcher has added information based on insights from the literature review of previous steps as the presentation to other researchers.

Fundamentals:

<u>Team elements:</u>

1. **Hierarchy:**
   DataOps has no hierarchy. Together, teams are self-organized but fully and equally responsible for the results of deliverables.
2. **Team size:**
   Teams are between 4-12 persons, depending on the requirements and complexity. A minimum of 4 is required for having at least two team members with data-focused expertise

and two members with ops-focused expertise. Architect(s) or DataOps Mentor(s) are not considered as team member, but can consult the team.

3. **Members have their own expertise, and roles are flexible:**
   Within teams, specializations get valued. Individuals share their knowledge, automate tasks and explain their reasoning to colleagues. The methodology values learning from both end users/operations, in order to get requirements better specified, and data scientists, by integrating the received (technical) feedback.
4. **Be open and transparent:**
   Every choice is discussed, since they are seen as team choices. They get documented and are accessible inside and outside the team. Progression is traceable.
5. **Automate everything:**
   To have presented models react quickly on new data observations, DataOps has a high focus on automation. Is all effort done, teams have to guarantee they never do the exact same thing again. This is not equal to forbidding reconsidering things, what can be a result of several new observations that have such impact that fine-tuning the model will not be a fitting reaction.

Duration of process(steps):

6. **Members meet daily and deliver weekly:**
   A process step needs to get completed in 7 calendar days. During the week, the team meets daily for maximum 15 minutes to discuss progression in order to meet deadlines. Members can call for help. When having the weekly meeting, lasting for 60 minutes at max, the teams discuss work had been done and they set next week's planning. Work that will not be planned for next week, gets on the 'backlog' that gets ordered by phase.
7. **Discuss next week's work, including the relevance:**
   Teams discuss the to-do-list for the next loop period, lasting 7-days, including reviewing the relevance of planned work. Teams rather work on parts of tasks, rather than on month's lasting dead ending tasks. Teams avoid overcapacity and inefficiency by e.g. not cleaning thousands of records, but take a part of a dataset to do first tests and extend from there.
8. **Give value to first insights as a small step rather than waiting for full optimization**
   Just as using not all available data but a sample or a part of the dataset, adjusting the plan during the week is allowed. Extending the deadline is not. When the first results are delivered, a new loop can be made to improve the accuracy.
9. **Define the *why* before starting KD-efforts, but changing the *how* is always welcome**
   When the strategic need is clear for all parties, how you get there can be discovered underway.. When new insights lead to a need to get back to choices, iterations or adjustments can be made to get towards optimization.
10. **One iteration during a loop**
    When iterating, the team has another 7 days to work on a process step. By having maximum one iterating during a complete loop, the process avoids to stay stuck at one step and works towards first deliverables.

Quality and process principles:

11. **Learning:**
    Since the methodology is based on skilled professionals, teams quickly win in effectiveness by continuously learning from each other. By discussing daily and delivering weekly, members can continuously measure their progress in perspective of mutual expectations.
12. **Big data does enrich the model:**

The methodology seeks for a combination of information available from both within the organization as from external sources. A combination between both provides the best answer to requirements. Since automation is part of the methodology, data from both sources can get processed near real-time and deliver knowledge to end-users accordingly.

13. **Try to destroy before you deploy:**
    Before delivering the results teams need to do the analysis whether this model adds more value and potential risk. If not, improvements are obviously needed. Part of automating everything is bringing in rules that inform the team when delivered work goes beyond set boundaries. "What if…?" questions set the knowledge up for current and future success.
14. **Retrospective (process) and maintenance:**
    As a continuous subject of discussion, the process gets discussed to provide room for improvement where needed. Since the model welcomes Big Data, it provides data to the model with a potential need to deployed models and so, maintenance wins in importance. Faster changing requirements have the same impact. Since Data and Ops are well connected at any time, teams work on maintenance constantly.

Roles:

- **Data-focused members:**
  These members have their expertise mainly on the data-part of the activities within the KD-methodology.
- **Ops-focused members:**
  These members have a great understanding on what information is needed and how the business can benefit from the knowledge discovered.
- **Architect and DataOps Mentor:**
  As both are not members of the team, both have a consulting role for the teams. The architect mainly helps to create a connection between not only the goals of operations but also the strategy, processes and IT governance of the organization. The DataOps Mentor is a role that can get consulted to have the teams walk through the process steps effectively. Since the teams are fully responsible for their results, consultation in both roles is a team responsibility as well. Their success depends on the results of all DataOps teams together.

4.3 Design steps

New Business Request: A defined starting point
In the literature review, the paradigm of knowledge discovery as a continuous activity rather than a process with defined finishing step is provided. Building further on this paradigm, DataOps lacks a finishing point as well. Obviously, the loop starts with a "business problem" or question.

Problem understanding
A new request can be seen as a need, in the model presented as a problem based on the fact that the need is not easily fulfilled. By selecting stakeholders involved with that need, gives the opportunity is presented to align them in order to get one defined problem and (a set of) objective(s) and goal(s) for the KD-efforts.

Team members have to get a clear view about the 'why' of KD-efforts. Obviously, answers can vary from 'we just need that figure' to 'this contributes enormous – today and in following years'. Seeing the objective(s) and goal(s) in the business perspective provides the opportunity to transfer data into information efficiently and effectively.

Business understanding
In this phase, we review the connections of the defined problem within the organization. This includes the process of defining stakeholders and sources of the data. Additionally, it defines the analytics capability along three dimensions: the analytical environment of the organization, the quality of the data on an organizational level and how individuals let their decisions be guided by analytics.

Iteration (1):
By getting a better view about the business and relationship between the business and the KD-goals, a need to come to a sharper problem statement is a plausible scenario since in the end, all stakeholders have to be aligned about the definition of the problem and how to work towards a solution. Continuous change on *how* is welcome at any time, all stakeholders need to get involved on *why* before starting the "Data-part" and stay aligned afterwards.

Data understanding
For the proposed model, KD-workers need to get a good understanding about the playing field, the quality of the data, to come from data to information.

Data selection and preparation
Having a clear overview of the available data, one can select, transform and clean the data.

Iteration (2):
During the selection and preparation, lessons learned can lead to a need for better understanding of the data and their relationship.

Modeling
Having several techniques available for the same defined problem, KD-workers will have to try several ways to generate conversion routines from data to information. Although modeling, just one step of the "Data-part", sounds technical, "Ops"-professionals can contribute to their counterparts by periodically providing input on relevance.

Testing
Obviously, the diversity of models bring different outcomes, not all of which are relevant. To optimize the accuracy of the model(s) selected for the next step, evaluation, a possibility to iterate is provided.

Iteration (3):
Insights lead to new insights, e.g. about the available data(set). Developing a better understanding about the data leads to better decisions. This iteration ensures an efficient way within the "Data"-part of the DataOps methodology.

Evaluation
KD-efforts are only relevant when they support business objectives. When not having the match between information needs and delivered information, efforts are not effective. To prevent a forced deployment with depressing results, one could iterate and redo the "Data-part" once more (iteration 4) or redo the problem understanding and move forwards from there.

Iteration (4):
Lessons learned from the evaluation phase help KD-workers get a better understanding of the need of the business. Since the needed information will likely change, data understanding needs to get redone in order to loop once more through the "Data"-part.

Iteration (5):
First results can be delivered after having done one loop. Evaluation will lead to a conclusion about the risks and benefits of the model presented. Having this beta-state gives the team a possibility to present and deliver without a need to have a forced deployment when the model has not yet achieved the required quality. This iteration is the only iteration that can be taken more than once.

Deployment
As with all the phases in the model, this phase especially is a duo-owned phase. By deploying the model, the business will probably have many questions, both practical and technical. The team member with expertise on the data-side can both respond and fetch ideas for further improvement or future projects.

New data observations
The model can be enriched by using new data observations. These new data observations contain both new information collected over time or and new information sources. After a day, 24 hours of new data can potentially be put into the model. When the work is automated, new results are delivered with high efficiency. An example can be the relation between employees during said day and the revenue generated.

New information sources or tables can provide new data observations as well. If a data supplier provides new data that has not been implemented in the model yet, the model can be enriched by testing the added value of this. If this adds value, this will make the model even more robust.

4.4 Plan for testing the design

During the empirical phase, the steps as presented in section 3.2.4 are involved. Based on the presentation to other scientists, the role of architect is added and the documentation (e.g. fundamentals, roles, process steps) that comes with the model of DataOps is redefined. After this, interviews were conducted by using the Delphi method. The Delphi method is conducted between 1950 and 1960 aiming to have "the most reliable consensus of a group of experts" (Okoli & Pawlowski, 2004). Since then, the Delphi method became a popular way of conducting empirical research in information systems to deal with specific problem types and outcome goals like prioritization or grading and concept- or framework development (Okoli & Pawlowski, 2004). A match between the objectives of the research, to create a consensus about strengths and weaknesses about the purposed KD methodology, is found.

Although the Delphi method can be done very efficient by e.g. email or web, the first round of expert interviews is in a personal 1-to-1 conversation. This provides the opportunity to give a good explanation on the purposed DataOps methodology and to make sure questions and definitions are clear. The next round of the Delphi method will be done via email.

Selecting the experts
Below profession/role are selected for running the Delphi method. They are selected based on their current professional role and which role they would logically fulfil within the DataOps model.

Profession/Role:	Main focus in the model:	Organization:
Product Owner	Operations	AAR
Engineers- and Data Analysts	Data- and Operations	AAR
Data Engineer	Data-side	AAR
Data Scientist	Data-side	Sogeti
Data Scientist	Data-side	Sogeti
Data Scientist	Data-side	Sogeti

Table 6: Overview of interviewees

Provided information and documentation
Before the first round of interviews, interviewees have received:

- A short introduction on what to expect;
- A letter of confidentiality stating what information will get used (and what not);

During the interviews, interviewees have received (documentation):

- Evaluation criteria to evaluate the current KD-methodology (to be filled in);
- The DataOps esign description, including:
    - A general description;
    - The key principles and roles;
    - The design steps;
- Evaluation criteria to evaluate the DataOps methodology (to be filled in).

5 Demonstration

The next step done within this research were the expert interviews. Feedback from the interviewees was gained on the fundamentals, the process steps and the model itself. The interviews gave context to the situation and when filling in the questionnaire, interviewees had the opportunity to further elucidate on specific criteria. Section 5 presents the evaluation and section 6 presents the results from the second (expert interviews) and third (feedback on conclusions) phase. The expert interviews (5.1) with experts contained 5 steps, which are:

1. Interview questions on the interviewee's experience with Agile KD;
2. Interview questions on the current way of doing KD;
3. A questionnaire (evaluation) about the current way of doing KD;
4. Interview questions on the proposed way of doing KD;
5. A questionnaire (evaluation) on the proposed way of doing KD.

This round has led to conclusions. Following the Delphi method, these conclusions are shared with the interviewees (5.2) in order to let them add on, or change, their original answers.

5.1 Interviews with experts

Setting of the interviews
Six interviews are conducted, equally distributed on two organizations. Except one Skype-interview, all interviews were at the case organization and in person. Interviewees were open, willing to share their vision and audio recording was allowed during all the conversations. The interviews took between 1 hour and 1.5 hours.

Plan and reality
The 5 steps as described above are slightly different from the originally planned steps, as described in section 3.2.4. Based on the presentation provided towards other researchers, an extra step was added in order to get to know the agile-experience of the interviewee. Since the original plan had two somehow similar steps, we could mere these in order to get above steps.

The interview phase had an ambitious start by approaching several KD-experts. A list of 12 interviewees willing to participate was made. Unfortunately, based on deadlines for this research, there was no possibility to have interviews with everyone. A choice has been made to do 6 in-depth interviews.

Based on the choice to focus on two case organizations, where a more random selection of experts was originally planned, the preferred equal distribution between data- and ops focused is not achieved since four interviewees are data-focused and two interviewees are ops-focused.

Adjustments on the case organization
The original plan was to do the research within one organization, the employer of the researcher. During the first part of the research, the conclusion was drawn that his colleagues are mostly inexperienced with agile. In order to let this research be valuable for agile experienced and inexperienced users, the choice is made to conduct interviews within two organizations.

Differences were expected between the two case organizations, in this research a consulting company and the researcher's employer. This can affect the vision of the interviewees on the current and proposed methodology. Although all interviewees are actively working with knowledge discovery, a difference is noticed between agile experienced and inexperienced interviewees. Next sections will zoom in on this.

During the interviews, when going into detail on Agile-specific subjects, agile experienced interviewees are more able to level and respond on interview questions. Responses gained form the other group are valuable, but some more explanation on definitions was needed.

5.2 Feedback on interview conclusions (Delphi method)

After the interviews, a consolidation has been made of all the given answers. This consolidation, including documentation as provided during the interviews, has been provided to the interviewees with the request to read through all answers in order to give their opinion once more. This research step has been explained during the interviews.

Despite the interviews being open and participants showing willingness to answer questions and provide feedback during the interviews, expectations on this part of the research are not met. Most interviewees replied with a confirmation of the summary, including complements on the content. This can mean that the interviewees recognize many touchpoints with their point of view on the methodology, or they took the easiest way. Some interviewees took the time to come with feedback, actually confirming what was already in the document.

Although meeting the requirements as presented earlier, the researcher definitely had higher expectations from on the input out of this cycle of the research.

6 Evaluation

During the interviews, we focused on the current KD methodology and the proposed methodology of DataOps. As described in section 5, feedback is both gained by conducting interview questions and questionnaires. The questionnaire helps to get a further understanding of the interviewee's opinion of the methodology. A Likert scale is used. This scale has ratings from very negative (- - -) to very positive (+ + +) and various options in between. To do an analysis on the results, the answers are reflecting scores as stated in Table 6. For example: if a respondent is highly satisfied (+ + +), the answer scores 3 points. If a second respondent is satisfied (+), the answer scores 1 point. On average, satisfaction scores 2 points. By doing so, averages and total scores can be distillated from the responses. The conclusions and observations stated below are all distillated from the expert interviews.

- - -	-3
- -	-2
-	-1
+	+1
+ +	+2
+ + +	+3

Table 7: scores

6.1 Comparing of the methodologies

Expert interviews are conducted within two case organizations, one that is agile minded and one that is not. A difference is recognized. Interviewees working with that last organization describe projects as a linear process, somehow chaotic and not focused. A lot of things are unclear, like roles and responsivities. The current way of working lacks structure or guiding methodology. The participants within a project feel the need to deliver one, somehow perfect, end product and most of the time projects end without even delivering a result, making the current way of working highly inefficient.

Bringing useful data into a chaotic process will automatically result in discussions and confusion. Due to unclearness of roles and responsibilities, a huge overlap of work done is recognized and by not having structure or a (weekly) coordination session, results get discussed often. Like discussion on the results, people may have discussions about (data) definitions. Although not part of the scope of this research, most likely a lack of structured methodology results in confusion about the results of work done. Within a linear process, individuals working on a specific process step do not get insights on what happens in other processes, nor do they get feedback on the work they delivered.

Interviewees working agile are more positive about their current methodology, which can be best described as a self-build methodology inspired by agile Scrum. They see value in iterations, periodic updates and a better connection between (agile) teams. Stakeholder involvement and team dynamics are aspects where both groups can further optimize. Since this is one of the values of DataOps, the proposed methodology shows good potential.

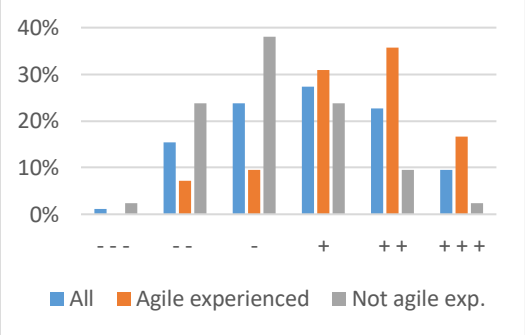


Figure 1; distribution of responses (current)

Attachment 7.1 and figure 2 show an overview on the distribution of all given answers. The average satisfaction with the current methodology is +0.43. Agile inexperienced interviewees are on average less satisfied (-0.79) than agile experienced interviewees (+1.19) and the difference between both groups is rather high (1.98) (see Attachment 7.9).

Responses on the proposed DataOps methodology were positive in general, combined with several useful advices and opportunities for further improvement, as presented later in this document. Several interviewees have requested to keep the documentation in order to use it for future projects. This can be seen as an indicator of satisfaction delivered by the structure from the methodology.

The DataOps methodology provides structure in order to work towards results in an effective way while the end-customer gets better involvement. The start and following steps are evaluated as logic, which also applies to the connections. It is mainly useful for bigger projects where several professionals are involved and documentation of work done is an important aspect. Frequent delivery is guaranteed, mainly by the weekly meetings.

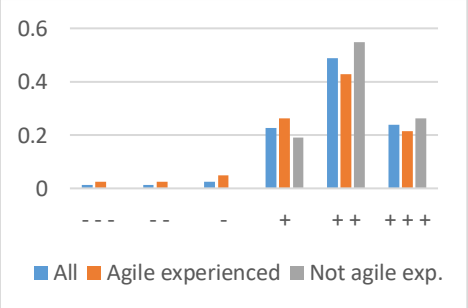


Figure 2: distribution of responses (proposed)

Attachment 7.3, 7.4 and figure 3 show the distribution of the provided answers on DataOps. If we follow the same analysis as we did for the starting situation, as shown in Attachment 7.10, the average evaluation of the criteria is +1.83 (was +0.43). Improvements were mainly noticed within the agile inexperienced group, as presented in

Attachment 7.10. Comparing the difference between average satisfaction levels for both groups in both situations, being the current and proposed methodology, this difference reduced from 1.98 to 0.45. This indicates a widely supported satisfaction.

6.2 Feedback categorized by evaluation criteria

Above, a general conclusion about the satisfaction is provided. When we look deeper into the evaluation criteria, we can bring to light the strengths and weaknesses of the methodology. Answers are based on the responses from the interviewees. Numbers behind the different criteria indicate the (average) shift of satisfaction between the current methodology and the proposed methodology.

1. Usable (from +1.17 to +2.17)
The cycle has all roles in it, which makes it very usable for big projects. It provides structure. Since there is the possibility to skip the deployment step, the methodology makes sure the product really adds value before implementing it in the business. Fundamentals help to get everyone 'on board'.

2. Efficient (from -0.50 to +1.50)
The methodology has predefined steps and rules about how to proceed during the process. The structure is mandatory and some process steps can be taken fast.

3. Reliable (from 0 to +1.83)
The biggest added value of the methodology concerning this criteria is the clearness of process steps and deliverables. Reliability on the output depends on the model and a human factor, which is part of the exploratory process of discovering new knowledge.

4. Maintainable (from +0.33 to +2.17)
The result of a project gets influenced in a positive way based on its maintainability, which is influenced by the way how a project is documented. Although this means some more work, it is easier to redo steps or to iterate earlier decisions.

5. Flexible (from +1.50 to +0.67)
This criteria requires some attention, since satisfaction drops where all other criteria win in satisfaction. The methodology brings teams 'stuck in a structure', but with some space via the possibility to iterate. Although it helps teams to get rid of 'freewheeling' situations, sometimes it works to have maximum flexibility.

6. Reusable (from +0.17 to +2.00)
Due to good documentation and automation, the methodology is very reusable. It is easy to take some steps backe and one can use lessons learned for other projects as well.

7. End-user involvement (from +0.33 to +2.00)
Some interviewees came from a waterfall-inspired situation, where communication is centered at the start of a project. Within the methodology, there is a continuous loop of sharing, which is evaluated as very positive. The assignment gets a better understanding when the end-user is involved. Although adjustments can be made, end-user requirements need to be as clear as possible when starting a project to avoid lacking an end-goal. Although the end user is a key concept of the methodology, the fundamentals have no bullet point giving the end user the required attention.

8. Frequent delivery (from +1.00 to +2.33)
Just as end-user involvement, the continuous communication during the project enforces that there is a continuous delivery. Several interviewees see this as a big improvement.

9. Guiding teamwork (from +0.50 to +2.17)
The methodology is clear on this aspect (roles, goals, team dynamics). Having frequent meetings ensures teamwork. When adjustments are needed, all relevant members are aware of what this means for their own work and the work of others. Based on team responsibility, the methodology lacks a 'manager' to make decisions when the team is stuck somewhere in the process. Having a senior within the team sometimes helps to keep things going into the right direction.

10. Iterations appreciated (from +1.17 to +1.33)
The methodology provides room to iterate, without the possibility to 'take to many steps back in the process'. The possibility to iterate just once avoids getting stuck in the process. When completing a loop, one can do the process again and take several process steps by just checking parameters. An explanation is needed about why we can iterate once per loop.

11. Fast delivery (from +0.50 to +1.50)
Interviewees expect a fast delivery cycle by taking out the possibility to keep working on one part of the process. There is a need to deliver. The fundamentals gave the impression that, when the product is not meeting expectations, the whole cycle has to be done again. This takes time. A further explanation on quickly doing process steps that have not changed a lot is needed in order to avoid the impression of the need to do all process steps again.

12. Respecting budget (hours/money) (from +0.17 to +1.67)
The methodology helps to make a good process plan, so there is an opportunity to set and monitor the budget. Since end-users are involved, they keep connected with the process and so with the budget.

13. Self-empowered teams (from -0.50 to +1.83)
Teams are fully responsible for the deliverables. Teams work closely together, both technical and operational oriented members. For several interviewees this is a big improvement. When teams work closely together and a good plan is in place, then the product will improve in quality.

The maximum number of 12 is seen a little high. By having 12 people in a team, working towards a consensus is hard. This is confirmed by Grady et al. (2018), where 5-9 people is defined as a good number for complementary skills and efficient team communication.

Furthermore, the model provides information on the Data and Ops side. However, the fundamentals provide no context on how they work together and how they learn from each other. This can be helpful to bridge the gap between both sides.

14. Stimulates a learning-curve (from +0.17 to +2.50)
Bringing people together is a 'formula for success'. Members will leave meetings with more understanding and knowledge compared to when they started. Having 'awareness' avoids having little islands of members doing their own part of the process. Somehow, there is a natural need to understand what colleagues are working on.

6.3 Summary of feedback from the expert interviews

By discussing the evaluation criteria during the expert interviews, several good results are recognized. On the methodology itself and 13 out of 14 evaluation criteria, interviewees expect the methodology will provide improvements. The criteria of flexibility show some attention. Because of the guidance and sometimes strictness of the methodology, some interviewees feel there may be a situation where more flexibility is required.

The expert interviews were conducted based on a hypothetical situation. Nevertheless, interviewees expect that especially the usability, reliability, maintainability, end-user involvement, frequent delivery, teamwork and self-empowered teams will improve both quality and satisfaction. Respecting budget as a criteria shows potential for improvement as well. These improvements give a good potential to solve the high failure rates, mainly coming from changing requirements, high costs, long lead-times.

As well, we have found opportunities for improvements, as described in some evaluation criteria. Some opportunities for improvement to highlight are:

- The end user and communication, collaboration and knowledge are currently not part of the fundamentals;
- It was not clear that, when the loop is done once more, process steps with a similar input don't have to be done once more;
- Based on one interviewee, the name "deployment" is a very short description of the process step. Within this step, as explained in the design steps, activities are focused on both *deployment and evaluation* - which would be a better name for this process step;
- It was not clear what was meant by describing new data observations. The explanation during the interviews led to a better understanding. An adjustment to the documentation is needed to provide that clearness in the first place;
- There is currently no step that asks teams to agree on a budget and monitor this budget during the project;
- The methodology does not explain how Data and Ops work together;

Next to that, the DataOps methodology needs to be valuable for all users, both agile experienced and not agile experienced. As described in section 5, during the interviews, agile inexperienced interviewees needed some explanation, e.g. on definitions. This has led to a need to get the fundamentals and process steps more easily to read.

6.4 Feedback on interview conclusions (Delphi method)

As presented in section 5.2, interviewees mainly confirmed the documentation they have received. Unfortunately, this cycle did not give any other insights or adjustments to make the methodology even more robust.

6.5 Evaluation summary

In general, the methodology adds value in comparison with the Knowledge Discovery process the interviewees currently used. We separated two groups based on their way of working. The group that is currently not working agile show relative big improvements, which is an indicator of the added value of agile in general. The agile experienced group show improvements as well. Specifically based on their feedback and the request form 2 out of 3 interviewees having the documentation of DataOps show the potential of the methodology as an extension to currently used agile methodologies. Both aspects are confirmed by doing statistics on the evaluation criteria.

Based on the openness of the interviewees, new opportunities for improvement became known. Adjustments are made to the methodology as presented in a potentially improved way in Attachment 8.

Iterating is not only an aspect that is important within the model, but as well for the methodology itself in order to adjust on several aspects. Next to those aspects, we can conclude based on two

rounds of the Delphi method that interviewees have evaluated the DataOps methodology as an improvement with good potential.

Improvements based on feedback
Based on valuable feedback from the interviews, above adjustments have resulted in an improved version of the DataOps methodology. The new DataOps model is presented in Attachment 8, just as the fundamentals and design steps.

7 Discussion, conclusions and recommendations

7.1 Conclusions

Although popularity of big data and analytics has attracted attention from both researchers and practitioners, statistics show high failure rates of Data Science projects. Based on a solid theoretical background, a new methodology had been collected. Information on changing requirements for dealing with data due to changing data characteristics and stakeholders' dynamics resulted in the DataOps methodology. DataOps extends the improvements that came from agile methodologies, trying to break down silos between build and run, or between Data and Ops, by bringing together specialist from both areas. The methodology contains fundamentals, process steps and the model itself.

The empirical part show promising results. To avoid bringing just a solution for a part of the problem, this research zoomed in on the methodology itself and a wide spectrum of 14 separate evaluation criteria. On both the methodology itself and 13 out of 14 evaluation criteria, interviewees provided great feedback (see section 6.2 and 6.3) sometimes including advice for further improvements of the methodology. Main adjustments are:

- Specific focus on the end-user within the documentation;
- More explanation is provided how Data and Ops work closely together;
- Specific focus on communication, collaboration and knowledge sharing within the documentation;
- A even sharper description of the process steps and how to handle when more or less time is required;
- Further explanation on definitions is added;
- Now, agreeing on the budget has become part of the process.

During the demonstration, the methodology itself is evaluated as a great solution and a big step forwards. If tested in a real project, interviewees expect that especially the usability, reliability, maintainability, end-user involvement, frequent delivery, teamwork and self-empowered teams will show big improvements. These improvements give a good potential to solve the high failure rates, mainly coming from changing requirements, high costs, long lead-times. Respecting budget as a criteria shows potential for improvement as well. We expect the criteria of flexibility is a point of attention for further research. Adjustments on the fundamentals and process steps, which help to better explain the flexibility a team has. This research adds to existing literature since no research is found that scientifically shows the potential of a similar methodology.

7.3 Practical implications

As presented in the introduction, the attention towards Knowledge Discovery is growing fast. Organizations spend more time and effort in KD-activities. Still, 85% of all KD projects fail and current methodologies don't provide a proper solution (J. Saltz et al., 2018). For this research, high

failure rates are seen as the main problem. The main question is how agile methodologies can contribute to the knowledge discovery process within organizations. This research shows the added value of agile in general and provides a methodology that potentially fills the need for a suitable solution. The proposed methodology focuses on Knowledge Discovery within teams working in organizations. For KD practitioners, the DataOps methodology can support all stakeholders within the KD-cycle to work together more closely in order to achieve goals. When executed correctly, Knowledge Discovery brings organizations an advantage on their competition.

7.4 Recommendations for further research

Although agility, as counterpart of a waterfall way of working, can be seen as accelerator for Software Development, Hemon et al. (2018) describes that agility has not broken down silos between build and run. Although DevOps is an extension of agility, prior literature provides little guidance on the scientific value of DevOps in bringing a structure in doing KD projects. Next to that, previous work on agile methodologies mainly describes process steps and activities, having little attention to team dynamics. DevOps introduces a focus on communication, collaboration and knowledge sharing between developers and operators. By researching the benefit of DevOps and researching existing methodologies in order to learn from best practices, this research extends existing literature by providing the DataOps methodology.

The methodology serves as a starting point for further research. The methodology has been tested by doing an artificial evaluation. The artificial evaluation is found effective to evaluate the methodology's scientific foundation and potential. The naturalistic evaluation derives strength in its validity by evaluating the performance of the methodology in a real environment against its original purpose of being a fitting solution to said problem (Venable et al., 2014). Interviews are conducted after presenting and discussing the methodology. A next step would be to test the methodology by doing a real project within an organization. An A/B test can prove the potential in comparison with other methodologies.

7.5 Reflection and limitations

The study is not without limitations. Most limitations are due to limited time, leaving room for future research on the subject of DevOps in Knowledge Discovery.

Within this research, feedback gained from the demonstration is based on a hypothetical project. Testing the methodology can be fruitful to get detailed feedback from professionals and can bring additional factors to light, leading to a needs to adjust the methodology.

Details of the empirical evaluation are available and the research approach is transparent, so other researchers can draw their own conclusions. When using statistics in the evaluation or when extracting data from interview responses, a larger number of interviews would be beneficial in order to predict the value of the DataOps methodology even better. The unequal distribution between Data- and Ops-focused interviewees is not seen as a limitation since differences between these interviewees' responses marginal. They provided comparable feedback on the methodology.

After a couple of interviews, the quality of the interviews itself grew recognizing a learning curve towards getting useful information. In particular, zooming in on the story behind the evaluation criteria adds value in that perspective. If I had the opportunity to further research this topic, I would use audio again, probably combined with video. Expressions can help to extract information from given answers, just as a detailed transcription of the interviews does.

Another limitation is that this research does separate agile experienced and agile inexperienced professionals, but does not look at other aspects as organizational, project characteristics or others. Future researchers may further investigate the relationship between these characteristics and the effectiveness of a methodology.

After searching for literature on the subject, we discovered new and relevant definitions for the search string. Although these could lead to new articles, the current literature review provided a good foundation for creating the methodology.

We started the Delphi method ambitious, thinking it is good way to get consensus on the strengths and weaknesses of the methodology. Although the first round, being the expert interviews, gave a lot of input, the expected feedback on the consolidation of the expert interviews was disappointing. If doing the research again, I probably would bring the interviewees together for a face-to-face discussion.

8. References

Abrahamsson, P., Conboy, K., & Wang, X. (2009). 'Lots done, more to do': the current state of agile systems development research. *European Journal of Information Systems*, *18*(4), 281–284. https://doi.org/10.1057/ejis.2009.27

Alnoukari, M., Alzoabi, Z., & Hanna, S. (2008). Applying Adaptive Software Development (ASD) agile modeling on predictive data mining applications: ASD-DM methodology. *Proceedings - International Symposium on Information Technology 2008, ITSim*, *2*(September 2008). https://doi.org/10.1109/ITSIM.2008.4631695

Alnoukari, M., & El Sheikh, A. R. A. (2012). *Knowledge Discovery Process Models: From Traditional to Agile Modeling. Business Intelligence and Agile Methodologies for KnowledgeBased Organizations CrossDisciplinary Applications*. Retrieved from http://www.igi-global.com/chapter/business-intelligence-agile-methodologies-knowledge/58566

Azevedo, A., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a Parallel Overview. *Proceeding of the IADIS European Conference on Data Mining 2008*, (January 2008), 182–185. https://doi.org/ISBN: 978-972-8924-63-8

Babb, J., Nørbjerg, J., Yates, D. J., & Waguespack, L. J. (2017). The Empire Strikes Back : The end of Agile as we know it ?, *8*(8).

Balle, A. R., Oliveira, M., Curado, C., & Nodari, F. (2018). How do knowledge cycles happen in software development methodologies? *Industrial and Commercial Training*, *50*(7–8), 380–392. https://doi.org/10.1108/ICT-04-2018-0037

Beck, K., Beedle, M., Bennekum, A. van, Cockburn, A., Cunningham, W., Fowler, M., … Thomas, D. (2001). Manifesto for agile software development. Retrieved from http://agilemanifesto.org/

Ben Ayed, M., Ltifi, H., Kolski, C., & Alimi, A. M. (2010). A user-centered approach for the design and implementation of KDD-based DSS: A case study in the healthcare domain. *Decision Support Systems*, *50*(1), 64–78. https://doi.org/10.1016/j.dss.2010.07.003

Bergh, C., Benghiat, G., & Strod, E. (2019). *The DataOps Cookbook*.

Bowley, R. (2017). The Fastest-Growing Jobs in the U.S. Based on LinkedIn Data.

Brechner, E. (2015). *Agile Project Management with Kanban*.

Brynjolfsson, E., Hitt, L. M., & Kim, H. H. (2011). Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance? *Ssrn*. https://doi.org/10.2139/ssrn.1819486

Cao, L. (2015). Actionable knowledge discovery and delivery. *Advanced Information and Knowledge Processing*, *2*(9781447165507), 287–312. https://doi.org/10.1007/978-1-4471-6551-4_14

Cao, L. (2017). Data Science: A Comprehensive Overview LONGBING. *ACM Computing Surveys*, *50*(3), 1–42. https://doi.org/10.1145/3076253

Chen, C. L. P., & Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, *275*, 314–347. https://doi.org/10.1016/j.ins.2014.01.015

Cios, K., & Kurgan, L. (2003). 1 . Trends in Data Mining and Knowledge Discovery, (January 2013), 0–26. https://doi.org/10.1007/1-84628-183-0

Côrte-Real, N., Oliveira, T., & Ruivo, P. (2017). Assessing business value of Big Data Analytics in European firms. *Journal of Business Research*, *70*, 379–390. https://doi.org/10.1016/j.jbusres.2016.08.011

Das, M., Cui, R., Campbell, D. R., Agrawal, G., & Ramnath, R. (2015). Towards Methods for Systematic Research On Big Data.

Davenport, T. H. (2012). Data Scientist: The Sexiest Job of the 21st Century.

Davenport, T. H. (2014). How strategists use “big data” to support internal business decisions, discovery and production. *Strategy and Leadership*, *42*(4), 45–50. https://doi.org/10.1177/019263659107553207

Debois, P. (2011). Devops : A Software Revolution in the Making ?, *24*(8).

Di Orio, G., Cândido, G., & Barata, J. (2015). The Adapter module: A building block for Self-Learning Production Systems. *Robotics and Computer-Integrated Manufacturing*, *36*, 25–35. https://doi.org/10.1016/j.rcim.2014.12.007

Dingsøyr, T., Nerur, S., Balijepally, V., & Moe, N. B. (2012). A decade of agile methodologies: Towards explaining agile software development. *Journal of Systems and Software*, *85*(6), 1213–1221. https://doi.org/10.1016/j.jss.2012.02.033

do Nascimento, G. S., & de Oliveira, A. A. (2012). An Agile Knowledge Discovery in Databases Software Process, (c), 56–64. https://doi.org/10.1007/978-3-642-34679-8_6

do Nascimento, G. S., & de Oliveira, A. A. (2012). An Agile Knowledge Discovery in Databases Software Process, (c), 56–64. https://doi.org/10.1007/978-3-642-34679-8_6

Dragland, A. (2013). “Big Data, for better or worse: 90% of world’s data generated over last two years.” ScienceDaily. ScienceDaily, 22 May 2013. Retrieved from https://www.sciencedaily.com/releases/2013/05/130522085217.htm

Gao, J., Koronios, A., & Selle, S. (2015). Towards A Process View on Critical Success Factors in Big Data Analytics Projects. *Twenty-First Americas Conference on Information Systems*, 1–14.

Gartner. (2015). Gartner Says Business Intelligence and Analytics Leaders Must Focus on Mindsets and Culture to Kick Start Advanced Analytics. Retrieved from https://www.gartner.com/newsroom/id/3130017

Gil, D., & Song, I. (2016). Modeling and Management of Big Data : Challenges and opportunities. *Future Generation Computer Systems*, *63*, 96–99. https://doi.org/10.1016/j.future.2015.07.019

Goepfert, J., & Shirer, M. (2018). Revenues for Big Data and Business Analytics Solutions... Retrieved from https://www.idc.com/getdoc.jsp?containerId=prUS44215218

Google. (2018a). Google Trends. Retrieved from https://trends.google.com/trends/explore?date=2008-10-20 2018-11-20&q=big data,data science,data analytics

Google. (2018b). Google Trends (2). Retrieved from https://trends.google.com/trends/explore?date=all&geo=US&q=devops

Grady, N. W., Payne, J. A., & Parker, H. (2018). Agile big data analytics: AnalyticsOps for data science. *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017*, *2018–Janua*, 2331–2339. https://doi.org/10.1109/BigData.2017.8258187

Hemon, A., Monnier-Senicourt, L., & Rowe, F. (2018). Job Satisfaction Factors and Risks Perception : An embedded case study of DevOps and Agile Teams (pp. 1–17). Thirty ninth International Conference on Information Systems, San Francisco.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information, *28*(1), 75–105.

Highsmith, J. (2000). Retiring lifecycle dinosaurs: Using Adaptive Software Development to meet the

challenges of a high-speed, high-change environment. *Software Testing & Quality Engineering (STQE)*, (August), 22–28.

Ho, A. (2017). Beyond the Dataset : Understanding Sociotechnical Aspects of the Knowledge Discovery Process Among Modern Data Professionals.

Hofmann, M., & Tierney, B. (2009). Development of an Enhanced Generic Data Mining Life Cycle (DMLC). *The ITB Journal*, *10*(1), 50–71. https://doi.org/10.21427/D75R0B

Jabbari, R., & Ali, N. (2016). What is DevOps ?: A Systematic Mapping Study on Definitions and Practices What is DevOps ? A Systematic Mapping Study on Definitions and Practices, (May). https://doi.org/10.1145/2962695.2962707

Janssen, M., van der Voort, H., & Wahyudi, A. (2017). Factors influencing big data decision-making quality. *Journal of Business Research*, *70*, 338–345. https://doi.org/10.1016/j.jbusres.2016.08.007

Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013). Big Data : Issues and Challenges Moving Forward.

KDNuggets. (2014). Methodology Survey. Retrieved from https://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html

Kurgan, L. A., & Musilek, P. (2006). A survey of Knowledge Discovery and Data Mining process models. *Knowledge Engineering Review*, *21*(1), 1–24. https://doi.org/10.1017/S0269888906000737

Labrinidis, A., & Jagadish, H. V. (2012). Challenges and Opportunities with Big Data, 2032–2033.

Larson, D., & Chang, V. (2016). A review and future direction of agile, business intelligence, analytics and data science. *International Journal of Information Management*, *36*(5), 700–710. https://doi.org/10.1016/j.ijinfomgt.2016.04.013

Lei, H., Ganjeizadeh, F., Jayachandran, P. K., & Ozcan, P. (2017). A statistical analysis of the effects of Scrum and Kanban on software development projects. *Robotics and Computer Integrated Manufacturing*, *43*, 59–67. https://doi.org/10.1016/j.rcim.2015.12.001

Li, Y., Thomas, M. A., & Osei-Bryson, K. M. (2016). A snail shell process model for knowledge discovery via data analytics. *Decision Support Systems*, *91*, 1–12. https://doi.org/10.1016/j.dss.2016.07.003

Li, Y., Thomas, M. A., & Osei-Bryson, K. M. (2017). Ontology-based data mining model management for self-service knowledge discovery. *Information Systems Frontiers*, *19*(4), 925–943. https://doi.org/10.1007/s10796-016-9637-y

Lim, C., Kim, K. H., Kim, M. J., Heo, J. Y., Kim, K. J., & Maglio, P. P. (2018). From data to value: A nine-factor framework for data-based value creation in information-intensive services. *International Journal of Information Management*, *39*(January 2017), 121–135. https://doi.org/10.1016/j.ijinfomgt.2017.12.007

Mariscal, G., Marbán, Ó., & Fernández, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. *Knowledge Engineering Review*, *25*(2), 137–166. https://doi.org/10.1017/S0269888910000032

Matharu, G. S. (2015). Empirical Study of Agile Software Development Methodologies : A Comparative Analysis, *40*(1), 1–6. https://doi.org/10.1145/2693208.2693233

Matsudaira, K. (2015). The science of managing data science. *Communications of the ACM*, *58*(6),

44–47. https://doi.org/10.1145/2745390

Mikalef, P., Pappas, I. O., Krogstie, J., & Giannakos, M. (2018). Big data analytics capabilities: a systematic literature review and research agenda. *Information Systems and E-Business Management*, *16*(3), 547–578. https://doi.org/10.1007/s10257-017-0362-y

Miller, H. G., & Mork, P. (2013). From data to decisions: A value chain for big data. *IT Professional*, *15*(1), 57–59. https://doi.org/10.1109/MITP.2013.11

Mishra, A., Garbajosa, J., Wang, X., & Bosch, J. (2017). Future directions in Agile research : Alignment and divergence between research and practice : Agile Product Development Special Issue Of J . Softw . Future Directions in Agile Research : Alignments and Divergence between Research and Practice, (June). https://doi.org/10.1002/smr

Okoli, C., & Pawlowski, S. D. (2004). The Delphi method as a research tool : an example , design considerations and applications, *42*, 15–29. https://doi.org/10.1016/j.im.2003.11.002

Okoli, C., & Schabram, K. (2010). A Guide to Conducting a Systematic Literature Review of Information Systems Research, Sprouts: Working Papers on Information Systems, *10*(2010). https://doi.org/10.2139/ssrn.1954824

Oztemel, E., & Gursev, S. (2018). Literature review of Industry 4.0 and related technologies. *Journal of Intelligent Manufacturing*, (January).

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). High-pressure P-V relation and Grüneisen parameter for elemental strontium. https://doi.org/10.2753/MIS0742-1222240302

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2008). A Design Science Research Methodology for Information Systems Research, 45–77. https://doi.org/10.2753/MIS0742-1222240302

Piatetsky, G. (2018). How many data scientists are there and is there a shortage?

Pirttimäki, V., Lönnqvist, A., & Karjaluoto, A. (2005). Measurement of Business Intelligence in a Finnish Telecommunications Company. *Proceedings of the European Conference on Knowledge Management, ECKM*, *0530*, 451–458. https://doi.org/10.1201/1078.10580530/45769.23.1.20061201/91770.4

Popovič, A., Hackney, R., Coelho, P. S., & Jaklič, J. (2012). Towards business intelligence systems success: Effects of maturity and culture on analytical decision making. *Decision Support Systems*, *54*(1), 729–739. https://doi.org/10.1016/j.dss.2012.08.017

Saltz, J. (2015). The need for new processes, methodologies and tools to support big data teams and improve big data project effectiveness. *Proceedings - 2015 IEEE International Conference on Big Data, IEEE Big Data 2015*, 2066–2071. https://doi.org/10.1109/BigData.2015.7363988

Saltz, J., & Heckman, R. (2018). A Scalable Methodology to Guide Student Teams Executing Computing Projects. *ACM Transactions on Computing Education*, *18*(2), 1–19. https://doi.org/10.1145/3145477

Saltz, J., Hotz, N., Wild, D., & Stirling, K. (2018). Exploring project management methodologies used within data science teams. *24th Americas Conference on Information Systems 2018: Digital Disruption, AMCIS 2018*, 1–5. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-85054242714&partnerID=40&md5=14e13aa859562d5960f40bd528a5ef09

Saltz, J. S., Shamshurin, I., & Crowston, K. (2017). Comparing Data Science Project Management

Methodologies via a Controlled Experiment. *Proceedings of the 50th Hawaii International Conference on System Sciences*, 1013–1022. Retrieved from http://hdl.handle.net/10125/41273

Sangupamba Mwilu, O., Comyn-Wattiau, I., & Prat, N. (2016). Design science research contribution to business intelligence in the cloud — A systematic literature review. *Future Generation Computer Systems*, *63*(2016), 108–122. https://doi.org/10.1016/j.future.2015.11.014

Saunders, M., Lewis, P., & Thornhill, A. (2016). *Research methods for business students*. Harlow: Pearson.

Schmidt, C., & Nan Sun, W. (2018). Synthesizing agile and knowledge discovery: case study results. *Journal of Computer Information Systems*, *58*(2), 142–150.

Sharma, S. (2016). Expanded cloud plumes hiding Big Data ecosystem. *Future Generation Computer Systems*, *59*, 63–92. https://doi.org/10.1016/j.future.2016.01.003

Sharma, S., & Osei-Bryson, K. M. (2009). Framework for formal implementation of the business understanding phase of data mining projects. *Expert Systems with Applications*, *36*(2 PART 2), 4114–4124. https://doi.org/10.1016/j.eswa.2008.03.021

Sharma, V., Stranieri, A., Ugon, J., Vamplew, P., & Martin, L. (2017). An Agile Group Aware Process beyond CRISP-DM. *Proceedings of the International Conference on Compute and Data Analysis - ICCDA '17*, 109–113. https://doi.org/10.1145/3093241.3093273

Sharma, V., Stranieri, A., Ugon, J., Vamplew, P., & Martin, L. (2017). An Agile Group Aware Process beyond CRISP-DM. *Proceedings of the International Conference on Compute and Data Analysis - ICCDA '17*, 109–113. https://doi.org/10.1145/3093241.3093273

Sharp, J. H., & Babb, J. S. (2018). Is Information Systems Late to the Party ? The Current State of DevOps Research in the Association for Information Systems eLibrary, 1–8.

Shen, B. (2018). Universal knowledge discovery from big data using combined dual-cycle. *International Journal of Machine Learning and Cybernetics*, *9*(1), 133–144. https://doi.org/10.1007/s13042-015-0376-z

Thomopoulos, T. (2018). 3th Predictive Analytics and Big Data Forum. Amsterdam - 2018/09/28: Google.

Venable, J., Pries-heje, J., & Baskerville, R. (2014). RESEARCH ESSAY FEDS : a Framework for Evaluation in Design Science Research, (August), 1–13. https://doi.org/10.1057/ejis.2014.36

Vidgen, R., Shaw, S., & Grant, D. B. (2017). Management challenges in creating value from business analytics. *European Journal of Operational Research*, *261*(2), 626–639. https://doi.org/10.1016/j.ejor.2017.02.023

Walker, J. (2017). Big data strategies disappoint with 85 percent failure rate. Retrieved from http://www.digitaljournal.com/tech-and-science/technology/big-data-strategies-disappoint-with-85-percent-failure-rate/article/508325

Watson, R. T., & Webster, J. (2002). Analysisng The Past to Prepare for The Future: Writing Literatur Review. *MIS Quarterly Vol. 26 No. 2, Pp. Xiii-Xxiii/June 2002*, *26*(2). https://doi.org/10.1.1.104.6570

Wilkes, S. (2012). Some {Impacts} of "{Big} {Data}" on {Usability} {Practice}. *Commun. Des. Q. Rev*, *13*(2), 25–32. https://doi.org/10.1145/2424840.2424842

Zeng, J., & Glaister, K. W. (2018). Value creation from big data: Looking inside the black box. *Strategic Organization*, *16*(2), 105–140. https://doi.org/10.1177/1476127017697510

Zhou, L., Pan, S., Wang, J., & Vasilakos, A. V. (2017). Machine learning on big data: Opportunities and challenges. *Neurocomputing*, *237*(January), 350–361. https://doi.org/10.1016/j.neucom.2017.01.026

Attachment 1a – literature obtained from literature review (query)

The following articles where obtained from the query as defined in section 2.3:

				61	53	38			
Art Nr.	**Article title (first 40 caracters)**	**Year**	**Name 1the writer**	**Avai-lable?**	**Saved?**	**Practi-cal?**	**Contribution to:**		
							RQ1	**RQ2**	**RQ3**
L1	A review and future direction of agile, ...	2016	Larson	Yes	Yes	Yes	X	X	0
L2	Ambidextrous organization and agility in...	2018	Rialti	Yes	Yes	Yes	0	0	0
L3	The Adapter module: A building block for...	2015	Di Orio	Yes	Yes	Yes	X	0	0
L4	Identifying nonconformity root causes us...	2015	Donauer	Yes	Yes	Yes	0	0	0
L5	Big data analytics in logistics and supp...	2016	Wang	Yes	Yes	No			
L6	On the model design of integrated intell...	2015	Chen	Yes	Yes	Yes	0	0	0
L7	About Our Authors	2012	Unknown - bias	Yes	No				
L8	A Scalable Methodology to Guide Student ...	2018	Saltz	Yes	Yes	Yes	X	0	X
L9	Strategic Prototyping for Developing Big...	2016	Chen	No					
L10	Strategic Prototyping for Developing Big...	2016	Chen	No					
L11	Designing a 'concept of operations' arch...	2016	Harrington	Yes	Yes	Yes	0	0	0
L12	Expanded cloud plumes hiding Big Data ec...	2016	S. Sharma	Yes	Yes	Yes	X	0	0
L13	Some impacts of "big data" on usability ...	2012	Wilkes	Yes	Yes	Yes	X	0	0
L14	Privacy-by-design in big data analytics	2014	Monreale	Yes	No				
L15	Ambidextrous organization and agility in...	2018	Rialti	Yes	No				
L16	Prioritising data items for business ana...	2016	Pape	Yes	Yes	Yes	0	0	0
L17	Developing software systems to Big Data ...	2017	Osvaldo	Yes	Yes	Yes	0	0	0
L18	Parallel data intensive applications usi...	2015	Han	Yes	Yes	Yes	0	0	0
L19	A characterization of workflow managemen...	2017	Ferreira da Silva	Yes	Yes	Yes	0	0	0
L20	The digital transformation of intelligen...	2017	Ashwell	Yes	Yes	Yes	0	0	0
L21	Lessons learned from applying social net...	2015	Teixeira	Yes	No				
L22	Genetic Programming for Automating the D...	2012	Archanjo	No					

L23	A snail shell process model for knowledg...	2016	Li	Yes	Yes	Yes	X	0	0
L24	A user-centered approach for the design ...	2010	Ben Ayed	Yes	Yes	Yes	0	X	0
L25	A human-centred design approach for deve...	2013	Ltifi	Yes	Yes	Yes	0	0	0
L26	Data mining and clustering in chemical p...	2018	Thomas	Yes	Yes	Yes	0	0	0
L27	The role of intelligent agents and data ...	2012	Warkentin	Yes	Yes	Yes	0	0	0
L28	Concurrent multiresponse non-linear scre...	2015	Besseris	Yes	No		0	0	0
L29	A Cognitive Adopted Framework for IoT Bi...	2015	Mishra	No					
L30	Automated extraction of fragments of Bay...	2017	Trovati	Yes	Yes	Yes	0	0	0
L31	An intelligent system on knowledge gener...	2018	Sermet	Yes	Yes	Yes	0	0	0
L32	Management challenges in creating value ...	2017	Vidgen	Yes	Yes	Yes	0	X	X
L33	Information and reformation in KM system...	2017	Intezari	Yes	Yes	Yes	0	0	0
L34	Text Mining in Organizational Research	2018	Kobayashi	Yes	No				
L35	A multi-agent based system with big data...	2016	Giannakis	Yes	Yes	Yes	0	0	0
L36	Single imputation with multilayer percep...	2015	Silva-Ramirez	Yes	Yes	No			
L37	Tool allocation to smooth work-in-proces...	2018	Chien	Yes	Yes	No			
L38	Capturing value from big data – a taxono...	2016	Hartmann	Yes	Yes	Yes	0	0	0
L39	Partner selection of agricultural produc...	2016	He	Yes	Yes	No			
L40	Toolkit to help HR directors recruit mor...	2018	?	Yes	No				
L41	A case based reasoning approach on suppl...	2011	Zhao	Yes	Yes	No			
L42	A flexible data acquisition system for s...	2018	Fernández-García	Yes	Yes	No			
L43	Exploiting smart e-Health gateways at th...	2018	Rahmani	Yes	Yes	No			
L44	A two-phase MILP approach to integrate o...	2018	Sadic	Yes	Yes	No			
L45	Group topic modeling for academic knowle...	2012	Daud	Yes	Yes	No			
L46	Strategic Foresight of Future B2B Custom...	2018	Gentner	Yes	Yes	Yes	0	0	0
L47	Defect prediction from static code featu...	2010	Menzies	Yes	Yes	No			
L48	Automated knowledge discovery and sema...	2016	Lin	Yes	Yes	Yes	0	0	0
L49	A framework for ontology based decision ...	2012	Bhattacharya	Yes	Yes	Yes	0	0	0
L50	A framework to explore innovation at SAP...	2015	James	Yes	Yes	No			
L51	Where are human subjects in Big Data res...	2016	Metcalf	Yes	Yes	No			

L52	Hyper-ellipsoidal clustering technique f...	2014	Rehman	Yes	Yes	Yes	0	0	0
L53	Formal model for assigning human resourc...	2011	André	Yes	Yes	Yes	0	0	0
L54	Improving workflow design by mining reus...	2015	Tosta	Yes	Yes	Yes	0	0	0
L55	Factors influencing big data decision-ma...	2017	Janssen	Yes	Yes	Yes	X	O	O
L56	Profiling effects in industrial data min...	2012	Besseris	Yes	Yes	No			
L57	Toward a novel art inspired incremental ...	2017	Cheraghchi	Yes	Yes	No			
L58	Implementing Smart Factory of Industrie ...	2016	Wang	Yes	Yes	Yes	0	0	0
L59	Research on conceptual modeling: Themes,...	2015	Storey	Yes	Yes	Yes	0	0	0
L60	Big data and business analytics ecosyste...	2011	Pappas	Yes	Yes	No			
L61	Distributed data mining for e-business	2014	Liu	Yes	Yes	Yes	0	0	0
L62	Developing a Semantic Web Model for Medi...	2014	Mohammed	Yes	No				
L63	Modeling and optimizing large-scale data...	2013	Wöhrer	Yes	Yes	Yes	0	0	0
L64	Mining affective needs of automotive ind...	2018	Mavridou	Yes	Yes	No			
L65	Emerging trends, issues and challenges i...	2018	Kobusinska	Yes	Yes	Yes	0	0	0
L66	Linking Data-Rich Environments with Serv...	2017	Troilo	Yes	Yes	Yes	0	0	0
L67	Mining event logs to support workflow re...	2012	Liu	Yes	Yes	No			
L68	On process model synthesis based on even...	2016	Mitsyuk	Yes	Yes	Yes	0	0	0
L69	A new paradigm for serious games: Transm...	2014	Raybourn	Yes	Yes	No			
L70	An AHP-based approach toward enterprise ...	2011	Razavi	Yes	Yes	Yes	0	0	0
L71	Running state of the high energy consumi...	2017	Zhao	Yes	Yes	No			
L72	Cloud-based Wireless Network: Virtualize...	2015	Chen	Yes	Yes	No			
L73	Placing Humans in the Feedback Loop of S...	2013	Müller	Yes	Yes	No			
L74	How do top- and bottom-performing compa...	2017	Cao	Yes	Yes	Yes	0	0	0
L75	Intelligent services for Big Data scienc...	2014	Dobre	Yes	Yes	No			
L76	Developing a business analytics methodol...	2018	Hindle	Yes	Yes	Yes	0	0	0
L77	Understanding the value of big data in s...	2018	Brinch	Yes	Yes	Yes	0	0	0
L78	Distinctive organisational business impe...	2017	Smede	Yes	Yes	No			
L79	Assessing business value of Big Data Ana...	2017	Côrte-Real	Yes	Yes	Yes	0	X	X
L80	Big data initiatives in retail environme...	2018	Aloysius	Yes	Yes	No			

L81	Future Internet technologies for environ...	2016	Granell	Yes	Yes	Yes	0	0	0
L82	GRASP-based resource re-optimization for...	2016	Palmieri	Yes	Yes	No			
L83	Sustainable operations modeling and data...	2018	Gunasekaran	Yes	Yes	No			
L84	SemLinker: automating big data integrati...	2018	Alrehamy	Yes	Yes	No			
L85	A Survey on NoSQL Stores	2018	Davoudian	Yes	Yes	No			
L86	Optimization technology in cloud manufac...	2018	Guo	Yes	Yes	No			
L87	Model-driven development of a person-cen...	2013	Lachenmaier	Yes	Yes	No			
L88	Optimization Design of Marine Collective...	2018	Zhao	No					
L89	A global exploration of Big Data in the ...	2016	Richey	Yes	Yes	No			
L90	Avoid being the Turkey: How big data ana...	2018	Van Rijmenam	Yes	Yes	Yes	0	0	0
L91	Shifting intra- and inter-organizational...	2017	Scuotto	Yes	Yes	No			
L92	Holonic and multi-agent technologies for...	2017	Thomas	Yes	Yes	No			
L93	Literature review of Industry 4.0 and re...	2018	Oztemel	Yes	Yes	Yes	X	0	0
L94	Big data analytics: transforming data to...	2017	Bumblauskas	Yes	Yes	Yes	0	0	0
L95	Big data analytics in E-commerce: a syst...	2016	Akter	Yes	Yes	No			
L96	Design science research contribution to ...	2016	Sangupamba	Yes	Yes	Yes	0	0	0
L97	IoTDeM: An IoT Big Data-oriented MapRedu...	2018	Lu	Yes	Yes	No			
L98	Implementation of an Intelligent Indoor ...	2018	Yang	Yes	Yes	No			
L99	SLA based healthcare big data analysis a...	2018	Sahoo	Yes	Yes	No			
L100	Identifying potentially disruptive trend...	2017	Dotsika	Yes	Yes	Yes	0	0	0
L101	Adaptive pairing of classifier and imput...	2016	Sim	Yes	Yes	No			
L102	Distributed online Temporal Fuzzy Concep...	2017	De Maio	Yes	Yes	No			
L103	A software reference architecture for se...	2017	Nadal	Yes	Yes	Yes	0	0	X
L104	Information system selection for a suppl...	2018	Samvedi	Yes	Yes	No			
L105	Study on the mode of intelligent chemica...	2016	Ji	Yes	Yes	No			
L106	Energy-efficient adaptive networked data...	2015	Cordeschi	Yes	Yes	No			
L107	A speculative parallel simulated anneali...	2018	Wang	Yes	Yes	No			
L108	Safety or no safety in numbers? Governme...	2015	Amankwah-Amoah	Yes	Yes	No			
L109	Performance Modelling and Analysis of So...	2016	Miao	Yes	Yes	No			

L110	IOTSim: A simulator for analysing IoT ap...	2017	Zeng	Yes	Yes	No			
L111	Big Data promises value: is hardware tec...	2015	Bhat	Yes	Yes	No			
L112	A distributed frequent itemset mining al...	2015	Zhang	Yes	Yes	No			
L113	Value creation from big data: Looking in...	2018	Zeng	Yes	Yes	No			
L114	How strategists use "big data" to suppor...	2014	Davenport	Yes	Yes	Yes	0	0	0
L115	Big data analytics capabilities: a syste...	2018	Mikalef	Yes	Yes	Yes	0	X	0
L116	Networking women translators in Spain (1...	2018	Romero López	Yes	No				
L117	Sharing big biomedical data.	2015	Toga	Yes	Yes	No			
L118	Smart HR 4.0 – how industry 4.0 is disru...	2018	Sivathanu	Yes	No				
L119	The golden age for popularizing big data...	2016	Lionel Ming-Shuan	Yes	Yes	No			
L120	Using Big Data to Improve Customer Exper...	2014	Spiess	Yes	Yes	No			
L121	CIRUS: an elastic cloud-based framework ...	2016	Pham	Yes	Yes	No			
L122	Detours on the Path to a European Big Da...	2017	Engels	Yes	Yes	No			
L123	A model-driven approach for the formal v...	2017	Marconi	Yes	Yes	No			
L124	Unicorn data scientist: the rarest of br...	2017	Baškarada	Yes	Yes	Yes	0	0	X
L125	D-Ocean: an unstructured data management...	2016	Zhuang	Yes	Yes	Yes	0	0	0
L126	Logistics 4.0 and emerging sustainable b...	2017	Strandhagen	Yes	Yes	No			
L127	Data Science as an Innovation Challenge:...	2018	Kayser	Yes	Yes	Yes	0	0	0
L128	The new marketing solutions that will dr...	2016	Grossberg	Yes	Yes	No			
L129	EXPERIENCE: Succeeding at Data Manageme...	2016	Aiken	Yes	Yes	Yes	0	0	0
L130	Actionable Knowledge As A Service (AKAAS...	2015	Depeige	Yes	Yes	Yes	0	0	0
L131	Big data needn't be a big headache: How ...	2012	Not stated	Yes	Yes	No			
L132	The science of managing data science	2015	Matsudaira	Yes	Yes	Yes	X	0	0
L133	Scalable SQL	2011	Rys	Yes	Yes	No			

Attachment 1b – literature provided by the University

The following articles where provided by the University. Some articles, marked with * in the column “Available?”, where found by using the search query (see Attachment1a).

Art Nr.	Article title (first 40 caracters)	Year	Name 1the writer	Avail-able?	Saved?	Practi-cal?	Contribution to: RQ1	RQ2	RQ3
P1	A snail shell process model for knowledg...	2016	Li	*			X	0	X
P2	A survey of data mining and knowledge di...	2010	Mariscal	Yes	Yes	Yes	X	O	X
P3	Agile KDD, An Agile Knowledge Discovery in ...	2012	Do Nascimento	Yes	Yes	Yes	X	O	X
P4	An Agile Group Aware Process beyond CRIS...	2017	Sharma	Yes	Yes	Yes	X	O	X
P5	Synthesizing Agile and Knowledge Discove...	2016	Schmidt	Yes	Yes	Yes	O	O	X
P6	Universal knowledge discovery from big d...	2018	Shen	Yes	Yes	Yes	X	O	X
P7	The need for new processes, methodologie...	2015	Saltz	Yes	Yes	Yes	X	X	O
P8	Exploring How Different Project Manageme...	2017	Saltz	Yes	Yes	Yes			
P9	Exploring Project Management Methodolog...	2018	Saltz	Yes	Yes	Yes	X	O	X
P10	Comparing Data Science Project Manageme...	2017	Saltz	Yes	Yes	Yes	X		
P11	Applying adaptive software development…	2008	Alnoukari	Yes	Yes	Yes	O	X	X

Attachment 1c – literature obtained from references of articles from query or provided by the University.

The following articles where used as reference within one of the articles obtained from the search query or provided by the Open University. The source (Src.) tells that source.

Art Nr.	Src.	Article title (first 40 caracters)	Year	Name 1the writer	Avai-lable?	Saved?	Practic al?	Contribution to:		
								RQ1	RQ2	RQ3
S1	L1	A decade of agile methodologies: Towar…	2009	Abrahamsson	Yes	Yes	Yes		X	
S2	P6	Actionable knowledge discovery and delivery	2012	Cao	Yes	Yes	Yes	X		
S3	P3	Manifesto of Agile Software Development	2001	Beck	Yes	Yes	Yes		X	
S4	P2	Framework for formal implementation of t...	2009	Sharma	Yes	Yes	Yes	X		
S5	P7	Towards A Process View on Critical Succe...	2015	Gao	Yes	Yes	Yes			
S6	L1	A decade of agile methodologies: Towards e…	2012	Dingsøyr	Yes	Yes	Yes		X	X
S7	P11	KDD, semma and CRISP-DM: a parallel over...	2008	Azevedo	Yes	Yes	Yes	X		
S8	P11	Development of an Enhanced Generic Data…	2009	Hofman	Yes	Yes	Yes	X		
S9	P2	A survey of Knowledge Discovery and Data...	2006	Kurgan	Yes	Yes	Yes	X		

Attachment 2a – Knowledge Discovery Methodologies (1/2)

Model	KDD Process	SEMMA	Six Sigma	Cabena et al.	Anand & Buchner
No of steps	9		5	5	8
Reference	Fayyad	SAS Institute	Harry & Schroeder	Cabena	Anand
Year	1996	1996	1996	1998	1998
Steps	1. Developing and Understanding of Application Domain	1. Sample	1. Define	1. Business Objective Determination	1. Human Resource Identification
					2. Problem Specification
	2. Creating a Target Data Set	2. Explore		2. Data Preparation	3. Data Prospecting
					4. Domain Knowledge Elicitation
	3. Data Cleaning and Projection	3. Modify			5. Methodology Identification
	4. Data Reduction and Pre-processing		2. Measure		6. Data Pre-processing
	5. Choosing the DM Task				
	6. Choosing the DM Algorithm				
	7. DM	4. Model	3. Analyze	3. DM	7. Pattern Discovery
	8. Interpreting Mined Pattern	5. Assess	4. Improve	4. Domain Knowledge Elicitation	8. Knowledge Post-processing
	9. Consolidating Discovered Knowledge		5. Control	5. Assimilation of Knowledge	

Attachment 2b – Knowledge Discovery Methodologies (2/2)

Model	KDLC	CRISP-DM	Cios et al.	The snail shell	Combined dual-cycle methodology	
No of steps	8	6	6	7	Two cycles of 5 = 10	
Reference	Lee & Kerschberg	Chapman	Cios	Li et al.	Shen	
Year	1998	2000	2003	2016	2018	
Steps				1. Problem Formulation		
	1. Define the objectives	1. Business Understanding	1. Understanding the Problem Domain	2. Business Understanding	1a. Big data accumulation and acquisition	1b. System abstraction and modelling
	2. Select Relevant Business Data	2. Data Understanding	2. Understanding the Data	3. Data Understanding	2a. Big data pre-processing	2b. Theoretical analysis and hypothesis proposition
	3. Data Quality Analysis					
	4. Clean and Transform Data	3. Data Pre-preparation	3. Preparation of the Data	4. Data Preparation	3a. Data-driven analytics	3b. Simulations, experiments and data analysis
	5. Data Mining	4. Modelling	4. DM	5. Modelling	4a. Universal Knowledge discovery	4b. The emergence of universal phenomenon
	6. Acquire Knowledge	5. Evaluation	5. Evaluation of the Discovered Knowledge	6. Evaluation	5a. Presentation and post-processing	5b. Putting forward UKN and further verification
	7. Evaluate Results	6. Deployment	6. Using the Discovered Knowledge	7. Deployment		
	8. Deploy Results or Re-iterate			8. Maintenance		

Attachment 3 – Comparison of different agile-inspired methodologies

From (Matharu, 2015)

Parameter	Scrum	Extreme Programming	Kanban	DevOps
Design principle	Complex design	Simplification of code & accommo-dation of unexpected changes	Limits the amount of work in progress and reduces waste	Teams fully responsible of end result
Nature of customer interaction	Not compulsorily on-site*	On-site customer interaction	Not compulsorily on-site*	On-site customer interaction
Design complexity	Complex design	Simple design*	Simple visual design*	
Project coordination	Scrum Master	XP Coach	Team work	Team work
Roles assigned	3 pre-defined roles: product owner, scrum master & dev. team	No prescribed roles*	No prescribed roles*	Assigned professionals
Process ownership	Scrum master	Team ownership*	Team ownership*	Team ownership
Product Ownership	Product owner	Group responsibility *	Group responsibility*	Group responsibility
Team collaboration	Cross functional teams	Self-organizing teams	Team of specialized members	Self-organizing team of specialized members
Work flow approach	Iterations (sprints)	No iterations. Task flow	Short iterations	Iterations part of design
Requirements management	Requirements managed in form of artefacts through Sprint Backlog & product Backlog	Managed in form of Story Cards	Managed using Kanban Boards	Wide variety of stake-holders is involved in the project. Changing requirements are welcome!
Product delivery	Delivery as per Time Boxed Sprints	Continuous delivery*	Continuous delivery*	Continuous delivery, automated where possible
Coding standards	No coding standards*	Coding standards are used	No coding standards*	No coding standards*
Testing approach	No formal approach used for testing	Test driven development, including acceptance testing	Testing done after implementation of each work product	Continuously, for both ideas and directions as the end-product. Full involvement.
Accomodation of changes	Changes not allowed in sprints	Amenable to change even in later stages of development*	Changes allowed at any time*	Changes allowed at any time and welcomed by the team

Attachment 4 – Model Logbook

Date/period	Work done/Changes made
Sept. 2018 – Feb. 2019	Laying the literature foundation for the DataOps model.
Feb. 2019	Delivering the first three models of the model: - A basic version of the model; - The model including iterations; - The model including iterations and external influences.
Feb. 2019	Introduced a logic 'starting point'.
Mar. 2019	First iteration/review done using a group presentation with KD-experienced scientists. Model seems to have good potential. Based on feedback, the importance of maintenance is defined. The deployment phase has an iteration, bringing a possible "beta-state" to avoid the possibility to have a model with more risks then benefits. As well, the iterations (numbers) are put more logically and the explanation is enriched. Changed the order of the iterations. This is now logical.
Apr. 2019 / Mai 2019	The demonstration is done towards experts currently working within the field of Knowledge Discovery. Their feedback is presented in section 6.2. Based on received feedback, the model is adjusted as presented in Attachment9.
June 2019	The experts received a summary of the interviews and documentation (fundamentals and process steps) in order to give them the opportunity to provide feedback once again. From this round of the demonstration, no opportunities for improvement where received unfortunately.

Attachment 5 – Aspects of DataOps methodology and its source of inspiration

Source of inspiration: Key principles	
Team elements:	
Hierarchy	Agile manifesto, DevOps, Literature review (general)
Team setting and size	Agile manifesto, Scrum, DevOps, Literature review (general)
Expertise and roles	Agile manifesto, DevOps, Scrum
Open and transparent	Agile manifesto, DevOps
Automate everything	DevOps
Process elements:	
Meet daily, deliver weekly	Scrum, Kanban, Agile manifesto, DevOps,
Next week's to-do and relevance	Scrum, Kanban
Value first insights	Agile manifesto, Scrum, DevOps,
Welcome change on 'how'	Agile manifesto,
One iteration per loop	Own insight based on literature review (general)
Quality:	
Learning	DevOps,
Enrichment via big data	Dual-cycle methodology
Destroy before you deploy	KDLC, Dual-cycle methodology
Retrospective	The Snail Shell process model

Source of inspiration: Roles	
Data-focused members	Most methodologies, focusing just on these professionals
Ops-focused members	CRISP-DM (Missing stakeholder dynamics), Agile manifesto, DevOps,
Architect and Data Mentor	Scrum

Source of inspiration: Design steps	
A defined starting point	KDD
Problem understanding	Scrum, Snail Shell, CRISP-DM, Dual-cycle methodology
Business understanding	Scrum, Snail Shell, CRISP-DM
Data understanding	Snail Shell, CRISP-DM
Data selection and preparation	KDD, CRIPS-DM
Modeling	CRISP-DM
Testing	Agile manifesto, DevOps
Evaluation	CRISP-DM, Snail Shell
Deployment	KDD, CRISP-DM
External influences	Literature review (Big Data)
Maintenance	Snail Shell

Attachment 6 – Results on the criteria (individual interviews)

Interview A10

	- - -	- -	-	+	+ +	+ + +	
1. Usable					X X		**0**
2. Efficient			X X				**0**
3. Reliable		X			X		**+3**
4. Maintainable				X		X	**+2**
5. Flexible		X				X	**-4**
6. Reusable				X	X		**-1**
7. End-user involvement	X				X		**-4**
8. Frequent delivery						X X	**0**
9. Guiding teamwork				X	X		**+1**
10. Iterations appreciated				X		X	**-2**
11. Fast delivery			X			X	**+3**
12. Respecting budget (hours + budget €)				X		X	**+2**
13. Self-empowered teams		X			X		**+3**
14. Stimulates a learning-curve					X	X	**-1**
Overall increase/decrease							**+2**

Current KD process Green

Proposed KD process Orange

Interview A11

	- - -	- -	-	+	+ +	+ + +	
1. Usable					X X		**0**
2. Efficient				X	X		**+1**
3. Reliable				X	X		**-1**
4. Maintainable			X	X			**+1**
5. Flexible				X		X	**-2**
6. Reusable		X				X	**+4**
7. End-user involvement				X	X		**+1**
8. Frequent delivery				X	X		**+1**
9. Guiding teamwork				X X			**+1**
10. Iterations appreciated			X X				**0**
11. Fast delivery					X X		**0**
12. Respecting budget (hours + budget €)				X	X		**-1**
13. Self-empowered teams				X		X	**-2**
14. Stimulates a learning-curve				X		X	**+2**
Overall increase/decrease							**+5**

Current KD process Green

Proposed KD process Orange

Interview A12

	- - -	- -	-	+	+ +	+ + +	
1. Usable				X	X		**+1**
2. Efficient		X			X		**+3**
3. Reliable				X	X		**+1**
4. Maintainable			X		X		**+2**
5. Flexible		X			X		**+3**
6. Reusable		X			X		**+3**
7. End-user involvement			X		X		**+2**
8. Frequent delivery			X		X		**+2**
9. Guiding teamwork			X			X	**+3**
10. Iterations appreciated				X	X		**+1**
11. Fast delivery				X X			**0**
12. Respecting budget (hours + budget €)		X		X			**+2**
13. Self-empowered teams		X				X	**+4**
14. Stimulates a learning-curve			X		X		**+2**
Overall increase/decrease							**+29**

Current KD process Green

Proposed KD process Orange

Interview A13

	- - -	- -	-	+	+ +	+ + +	
1. Usable						X X	**0**
2. Efficient				X	X		**+1**
3. Reliable					X X		**0**
4. Maintainable					X	X	**+1**
5. Flexible				X X			**0**
6. Reusable				X X			**0**
7. End-user involvement					X X		**0**
8. Frequent delivery					X X		**0**
9. Guiding teamwork					X X		**0**
10. Iterations appreciated					X X		**0**
11. Fast delivery				X X			**0**
12. Respecting budget (hours + budget €)					X X		**0**
13. Self-empowered teams					X	X	**+1**
14. Stimulates a learning-curve					X X		**0**
Overall increase/decrease							**+3**

Current KD process Green

Proposed KD process Orange

Evaluation criteria: KD process *A14*

	- - -	- -	-	+	+ +	+ + +	
1. Usable				X	X		**+1**
2. Efficient			X		X		**+2**
3. Reliable			X	X			**+1**
4. Maintainable			X		X		**+2**
5. Flexible				X	X		**-1**
6. Reusable				X		X	**+2**
7. End-user involvement			X			X	**+3**
8. Frequent delivery		X				X	**+4**
9. Guiding teamwork					X X		**0**
10. Iterations appreciated					X X		**0**
11. Fast delivery		X		X			**+2**
12. Respecting budget (hours + budget €)			X		X		**+2**
13. Self-empowered teams			X		X		**+2**
14. Stimulates a learning-curve		X				X	**+4**
Overall increase/decrease							**+24**

Current KD process Green

Proposed KD process Orange

Evaluation criteria: KD process *A15*

	- - -	- -	-	+	+ +	+ + +	
1. Usable			X		X		**+2**
2. Efficient			X		X		**+2**
3. Reliable		X				X	**+4**
4. Maintainable					X X		**0**
5. Flexible				X	X		**-1**
6. Reusable				X	X		**+1**
7. End-user involvement			X			X	**+3**
8. Frequent delivery					X	X	-1
9. Guiding teamwork			X			X	**+3**
10. Iterations appreciated				X	X		**+1**
11. Fast delivery				X	X		**-1**
12. Respecting budget (hours + budget €)			X	X			**+1**
13. Self-empowered teams	X					X	**+5**
14. Stimulates a learning-curve		X				X	**+4**
Overall increase/decrease							**+23**

Current KD process Green

Proposed KD process Orange

Attachment 7 – Results on the criteria (individual interviews)

Attachment 7.1 - Ranking on the current KD methodology (all):

- - -	- -	-	+	+ +	+ + +
1%	15%	24%	27%	23%	10%

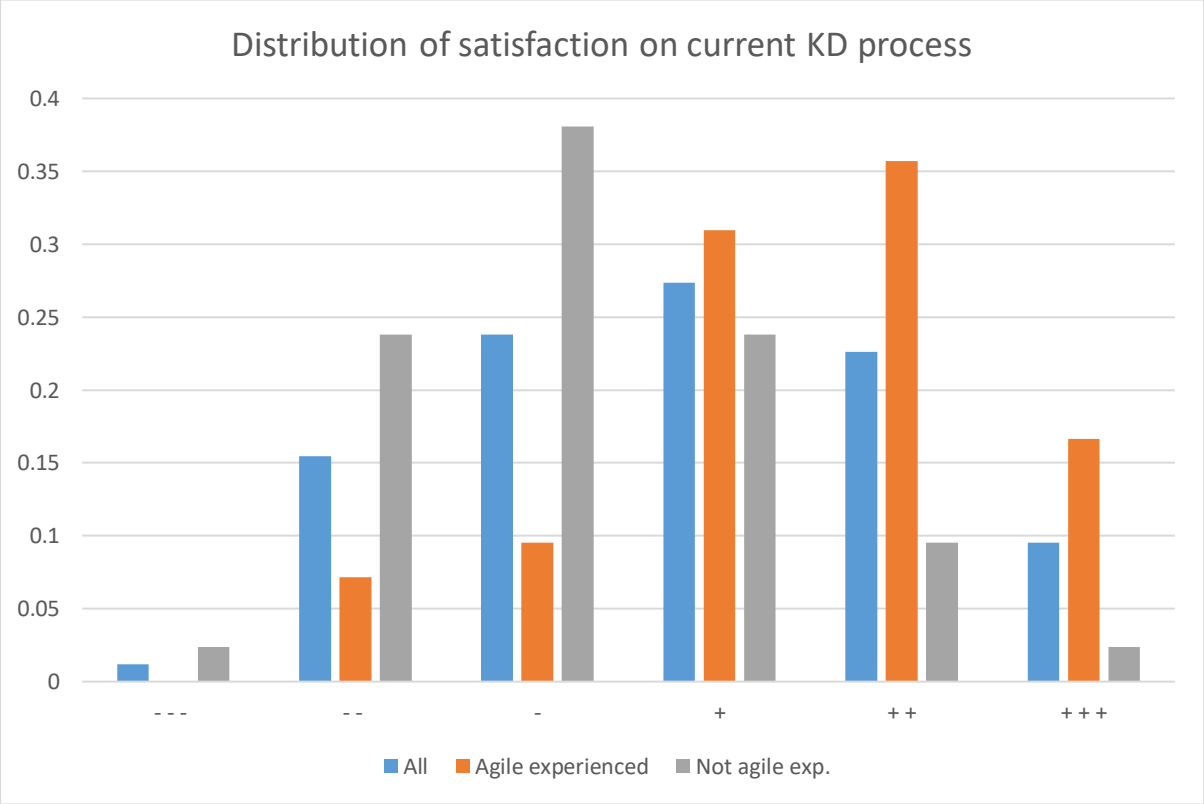
Attachment 7.2 - Ranking on the current KD methodology (agile and not agile compared):

Current KD methodology is agile (inspired)

- - -	- -	-	+	+ +	+ + +
0%	7%	10%	31%	36%	17%

Current KD methodology is not agile (inspired)

- - -	- -	-	+	+ +	+ + +
2%	24%	38%	24%	10%	2%



Attachment 7.3 - Ranking on the proposed KD methodology (all):

- - -	- -	-	+	+ +	+ + +
1%	1%	2%	23%	49%	24%

Attachment 7.4 - Ranking on the proposed KD methodology (agile and not agile compared):

Current KD methodology is agile (inspired)

- - -	- -	-	+	+ +	+ + +
2%	2%	5%	26%	43%	21%

Current KD methodology is not agile (inspired)

- - -	- -	-	+	+ +	+ + +
0%	0%	0%	19%	55%	26%

Attachment 7.5 - Difference (all):

- - -	- -	-	+	+ +	+ + +
0%	-14%	-21%	-5%	+ 26%	+ 14%

Attachment 7.6 – Difference in distribution the proposed KD methodology (agile and not agile compared):

Current KD methodology is agile (inspired)

- - -	- -	-	+	+ +	+ + +
2%	- 5%	- 5%	- 5%	+ 7%	+ 5%

Current KD methodology is not agile (inspired)

- - -	- -	-	+	+ +	+ + +
-2%	-24%	-38%	-5%	45%	24%

Attachment 7.7

Distribution of the evaluation criteria (starting situation - both agile and not agile)

	- - -	- -	-	+	+ +	+ + +
Usable	0%	0%	17%	50%	17%	17%
Efficient	0%	17%	50%	33%	0%	0%
Reliable	0%	33%	17%	17%	33%	0%
Maintainable	0%	0%	50%	17%	33%	0%
Flexible	0%	17%	0%	17%	33%	33%
Reusable	0%	33%	0%	50%	17%	0%
End-user involvement	0%	0%	50%	17%	33%	0%
Frequent delivery	0%	17%	17%	17%	17%	33%
Guiding teamwork	0%	0%	33%	50%	17%	0%
Iterations appreciated	0%	0%	17%	50%	17%	17%
Fast delivery	0%	17%	17%	33%	33%	0%
Respecting budget	0%	17%	33%	17%	33%	0%
Self-empowered teams	17%	33%	17%	0%	17%	17%
Stimulates a learning-curve	0%	33%	17%	17%	17%	17%

Attachment 7.8

Distribution of the evaluation criteria (proposed methodology - both agile and not agile)

	- - -	- -	-	+	+ +	+ + +
Usable	0%	0%	0%	0%	83%	17%
Efficient	0%	0%	17%	0%	83%	0%
Reliable	0%	0%	0%	33%	50%	17%
Maintainable	0%	0%	0%	17%	50%	33%
Flexible	0%	17%	0%	67%	17%	0%
Reusable	0%	0%	0%	33%	33%	33%
End-user involvement	17%	0%	0%	0%	50%	33%
Frequent delivery	0%	0%	0%	0%	67%	33%
Guiding teamwork	0%	0%	0%	17%	50%	33%
Iterations appreciated	0%	0%	17%	17%	67%	0%
Fast delivery	0%	0%	0%	67%	17%	17%
Respecting budget	0%	0%	0%	50%	33%	17%
Self-empowered teams	0%	0%	0%	17%	33%	50%
Stimulates a learning-curve	0%	0%	0%	0%	0%	50%

Attachment 7.9 – Ranking of the criteria for the current KD methodology

All interviewees

	- - -	- -	-	+	++	+++	
1. Usable	0	0	-1	3	2	3	**1.17**
2. Efficient	0	-2	-3	2	0	0	**-0.50**
3. Reliable	0	-4	-1	1	4	0	**0.00**
4. Maintainable	0	0	-3	1	4	0	**0.33**
5. Flexible	0	-2	0	1	4	6	**1.50**
6. Reusable	0	-4	0	3	2	0	**0.17**
7. End-user involvement	0	0	-3	1	4	0	**0.33**
8. Frequent delivery	0	-2	-1	1	2	6	**1.00**
9. Guiding teamwork	0	0	-2	3	2	0	**0.50**
10. Iterations appreciated	0	0	-1	3	2	3	**1.17**
11. Fast delivery	0	-2	-1	2	4	0	**0.50**
12. Respecting budget (hours + budget €)	0	-2	-2	1	4	0	**0.17**
13. Self-empowered teams	-3	-4	-1	0	2	3	**-0.50**
14. Stimulates a learning-curve	0	-4	-1	1	2	3	**0.17**
						ALL	**6.00**
						Average	**0.43**

Agile inspired

	- - -	- -	-	+	++	+++	
1. Usable	0	0	0	1	2	3	**2.00**
2. Efficient	0	0	-2	2	0	0	**0.00**
3. Reliable	0	-2	0	0	4	0	**0.67**
4. Maintainable	0	0	-2	1	2	0	**0.33**
5. Flexible	0	0	0	1	0	6	**2.33**
6. Reusable	0	-2	0	1	2	0	**0.33**
7. End-user involvement	0	0	0	1	4	0	**1.67**
8. Frequent delivery	0	0	0	1	2	3	**2.00**
9. Guiding teamwork	0	0	0	2	2	0	**1.33**
10. Iterations appreciated	0	0	-2	0	2	3	**1.00**
11. Fast delivery	0	0	-2	1	2	0	**0.33**
12. Respecting budget (hours + budget €)	0	0	0	1	4	0	**1.67**
13. Self-empowered teams	0	-2	0	0	2	3	**1.00**
14. Stimulates a learning-curve	0	0	0	1	2	3	**2.00**
						ALL	**16.67**
						Average	**1.19**

Not agile inspired

	- - -	- -	-	+	++	+++	
1. Usable	0	0	-1	2	0	0	**0.33**
2. Efficient	0	-2	-4	0	0	0	**-2.00**
3. Reliable	0	-2	-2	1	0	0	**-1.00**
4. Maintainable	0	0	-4	0	2	0	**-0.67**
5. Flexible	0	-2	0	0	4	0	**0.67**
6. Reusable	0	-2	0	2	0	0	**0.00**
7. End-user involvement	0	0	-6	0	0	0	**-2.00**
8. Frequent delivery	0	-2	-2	0	0	3	**-0.33**
9. Guiding teamwork	0	0	-4	1	0	0	**-1.00**
10. Iterations appreciated	0	0	0	3	0	0	**1.00**
11. Fast delivery	0	-2	0	1	2	0	**0.33**
12. Respecting budget (hours + budget €)	0	-2	-4	0	0	0	**-2.00**
13. Self-empowered teams	-3	-2	-2	0	0	0	**-2.33**
14. Stimulates a learning-curve	0	-4	-2	0	0	0	**-2.00**
						ALL	**-11.00**
						Average	**-0.79**

Attachment 7.10 – Ranking of the criteria for the proposed KD methodology

All interviewees

1. Usable	0	0	0	0	10	3	**2.17**
2. Efficient	0	0	-1	0	10	0	**1.50**
3. Reliable	0	0	0	2	6	3	**1.83**
4. Maintainable	0	0	0	1	6	6	**2.17**
5. Flexible	0	-2	0	4	2	0	**0.67**
6. Reusable	0	0	0	2	4	6	**2.00**
7. End-user involvement	0	0	0	0	6	6	**2.00**
8. Frequent delivery	0	0	0	0	8	6	**2.33**
9. Guiding teamwork	0	0	0	1	6	6	**2.17**
10. Iterations appreciated	0	0	-1	1	8	0	**1.33**
11. Fast delivery	0	0	0	4	2	3	**1.50**
12. Respecting budget (hours + budget €)	0	0	0	3	4	3	**1.67**
13. Self-empowered teams	-3	0	0	1	4	9	**1.83**
14. Stimulates a learning-curve	0	0	0	0	6	9	**2.50**
						ALL	**25.67**
						Average	**1.83**

Agile inspired

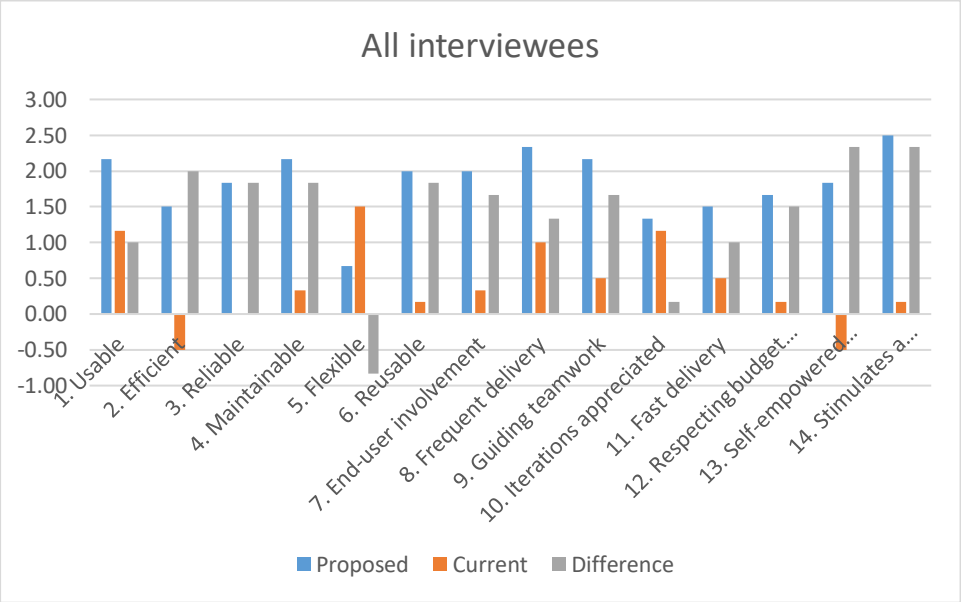
1. Usable	0	0	0	0	4	3	**2.33**
2. Efficient	0	0	-2	0	4	0	**0.67**
3. Reliable	0	0	0	1	4	0	**1.67**
4. Maintainable	0	0	0	1	0	6	**2.33**
5. Flexible	0	-2	0	2	0	0	**0.00**
6. Reusable	0	0	0	2	0	3	**1.67**
7. End-user involvement	0	0	0	0	4	0	**1.33**
8. Frequent delivery	0	0	0	0	4	3	**2.33**
9. Guiding teamwork	0	0	0	1	4	0	**1.67**
10. Iterations appreciated	0	0	-2	1	2	0	**0.33**
11. Fast delivery	0	0	0	1	2	3	**2.00**
12. Respecting budget (hours + budget €)	0	0	0	1	2	3	**2.00**
13. Self-empowered teams	-3	0	0	1	2	3	**1.00**
14. Stimulates a learning-curve	0	0	0	0	4	3	**2.33**
						ALL	**21.67**
						Average	**1.55**

Not agile inspired

1. Usable	0	0	0	0	6	0	**2.00**
2. Efficient	0	0	0	0	6	0	**2.00**
3. Reliable	0	0	0	1	2	3	**2.00**
4. Maintainable	0	0	0	0	6	0	**2.00**
5. Flexible	0	0	0	2	2	0	**1.33**
6. Reusable	0	0	0	0	4	3	**2.33**
7. End-user involvement	0	0	0	0	2	6	**2.67**
8. Frequent delivery	0	0	0	0	4	3	**2.33**
9. Guiding teamwork	0	0	0	0	2	6	**2.67**
10. Iterations appreciated	0	0	0	0	6	0	**2.00**
11. Fast delivery	0	0	0	3	0	0	**1.00**
12. Respecting budget (hours + budget €)	0	0	0	2	2	0	**1.33**
13. Self-empowered teams	-3	0	0	0	2	6	**1.67**
14. Stimulates a learning-curve	0	0	0	0	2	6	**2.67**
						ALL	**28.00**
						Average	**2.00**

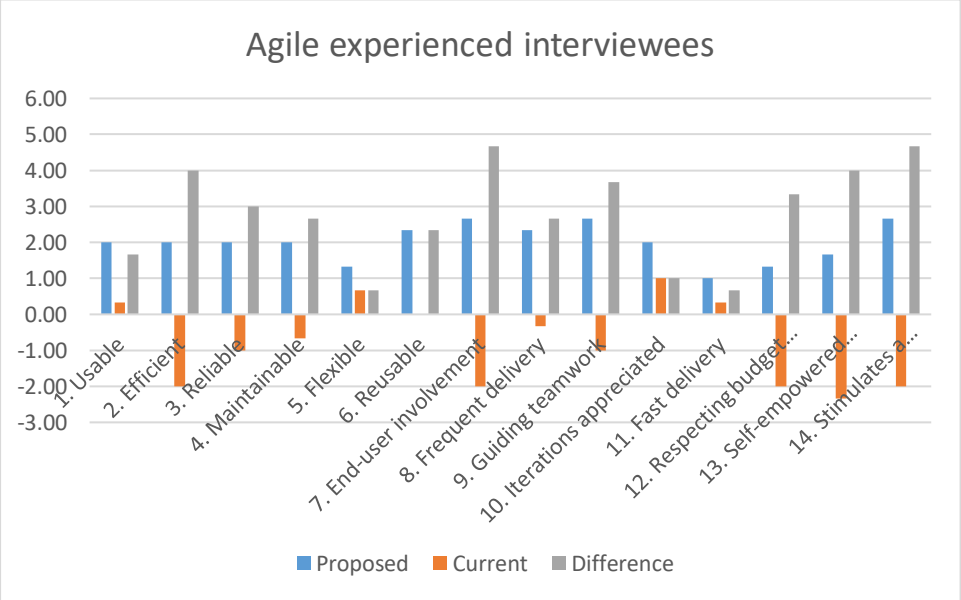
Attachment 7.11 – Differences when comparing individual evaluation criteria

	Current	Proposed	Difference
1. Usable	1.17	2.17	1.00
2. Efficient	-0.50	1.50	2.00
3. Reliable	0.00	1.83	1.83
4. Maintainable	0.33	2.17	1.83
5. Flexible	1.50	0.67	-0.83
6. Reusable	0.17	2.00	1.83
7. End-user involvement	0.33	2.00	1.67
8. Frequent delivery	1.00	2.33	1.33
9. Guiding teamwork	0.50	2.17	1.67
10. Iterations appreciated	1.17	1.33	0.17
11. Fast delivery	0.50	1.50	1.00
12. Respecting budget (hours + budget €)	0.17	1.67	1.50
13. Self-empowered teams	-0.50	1.83	2.33
14. Stimulates a learning-curve	0.17	2.50	2.33



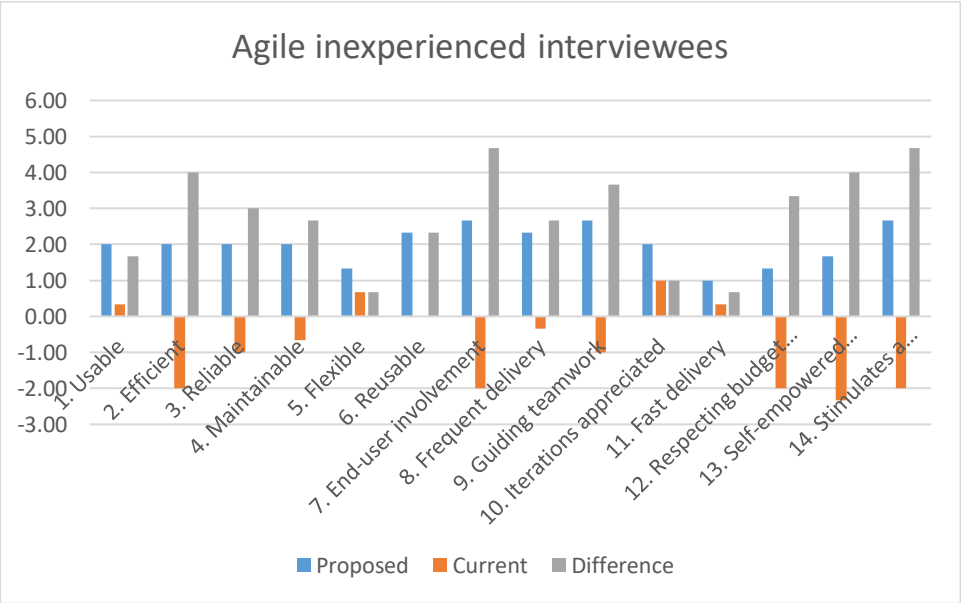
Agile inspired

	Current	Proposed	Difference
1. Usable	2.00	2.33	0.33
2. Efficient	0.00	0.67	0.67
3. Reliable	0.67	1.67	1.00
4. Maintainable	0.33	2.33	2.00
5. Flexible	2.33	0.00	-2.33
6. Reusable	0.33	1.67	1.33
7. End-user involvement	1.67	1.33	-0.33
8. Frequent delivery	2.00	2.33	0.33
9. Guiding teamwork	1.33	1.67	0.33
10. Iterations appreciated	1.00	0.33	-0.67
11. Fast delivery	0.33	2.00	1.67
12. Respecting budget (hours + budget €)	1.67	2.00	0.33
13. Self-empowered teams	1.00	2.00	1.00
14. Stimulates a learning-curve	2.00	2.33	0.33



Not agile inspired

	Current	Proposed	Difference
1. Usable	0.33	2.00	1.67
2. Efficient	-2.00	2.00	4.00
3. Reliable	-1.00	2.00	3.00
4. Maintainable	-0.67	2.00	2.67
5. Flexible	0.67	1.33	0.67
6. Reusable	0.00	2.33	2.33
7. End-user involvement	-2.00	2.67	4.67
8. Frequent delivery	-0.33	2.33	2.67
9. Guiding teamwork	-1.00	2.67	3.67
10. Iterations appreciated	1.00	2.00	1.00
11. Fast delivery	0.33	1.00	0.67
12. Respecting budget (hours + budget €)	-2.00	1.33	3.33
13. Self-empowered teams	-2.33	1.67	4.00
14. Stimulates a learning-curve	-2.00	2.67	4.67



Attachment 8 – Adjusted fundamentals, design steps and model

Adjusted fundamentals:

Team elements:

1. **Hierarchy:**
   DataOps has no hierarchy. Together, teams are self-organized but fully and equally responsible for the results of deliverables. Just in case when the team is stuck during the process, they can reach out to a senior in order to keep things going.
2. **Team size:**
   Teams are between 4-8 persons, depending on the requirements and complexity. A minimum of 4 is required for having at least two team members with data-focused expertise and two members with ops-focused expertise. They work together by explaining the work they have done and share results within the team. Communication, collaboration and knowledge sharing are important within teams working with DataOps. Architect(s) or DataOps Mentor(s) are not considered as team member, but can consult the team.
3. **Members have their own expertise, and roles are flexible:**
   Within teams, specializations get valued. Individuals share their knowledge, automate tasks and explain their reasoning to colleagues. The methodology values learning from both end users/operations, in order to get requirements better specified, and data scientists, by integrating the received (technical) feedback.
4. **Be open and transparent:**
   Every choice is discussed, since they are seen as team choices. They get documented and are accessible inside and outside the team. Progression is traceable.
5. **Automate everything:**
   To have presented models react quickly on new data observations, DataOps has a high focus on automation. New data observations can be both new data points and new datasets becoming available that can value to the model teams are building on. In all effort done, teams have to guarantee they never do the exact same thing again. This is not equal to forbidding reconsidering things, what can be a result of several new observations that have such impact that fine-tuning the model will not be a fitting reaction.

Duration of process(steps):

6. **Members meet daily and deliver weekly:**
   A process step needs to get completed in 7 calendar days. During the week, the team meets daily for maximum 15 minutes to discuss progression in order to meet deadlines. Members can call for help. When having the weekly meeting, lasting for 60 minutes at max, the teams discuss work had been done and they set next week's planning. Work that will not be planned for next week, gets on the 'backlog' that gets ordered by phase.
7. **Discuss next week's work with the team, including the end-user**
   Teams discuss the to-do-list for the next loop period, lasting seven days, including reviewing the relevance of planned work. Teams work on parts of tasks, rather than on month's lasting dead ending tasks. Teams avoid overcapacity and inefficiency by e.g. not cleaning thousands of records, but take a part of a dataset to do first tests and extend from there. The end-user stays involved by participating in these meetings in order to ensure the team keeps track of the right direction.
8. **Give value to first insights as a small step rather than waiting for full optimization**

Just as using not all available data but a sample or a part of the dataset, adjusting the plan during the week is allowed. Extending the deadline is not. When the first results are delivered, a new loop can be made to improve the accuracy.
The end user involvement is guaranteed not only in the beginning where teams set the requirements, but as well during the weekly meetings. In this way, they can track progression and let the teams make little adjustments to avoid delivering an end product that does not match expectations.

9. **Define the *why* before starting KD-efforts, but changing the *how* is always welcome**
   When the strategic need is clear for all parties, how you get there can be discovered underway. When new insights lead to a need to get back to choices, iterations or adjustments can be made to get towards optimization.
10. **One iteration during a loop**
    When iterating, the team has another seven days to work on a process step. By having maximum one iteration during a complete loop, the process avoids being stuck at one step and works towards first deliverables. If that first deliverable is ready, the team can redo the loop. Some steps need full attention of the team and some steps can be taken easily. If the parameters stay the same and the team agrees on that, a process step can be finished during the weekly meetings.

Quality and process principles:

11. **Learning:**
    Since the methodology is based on skilled professionals, teams quickly win in effectiveness by continuously learning from each other. By discussing daily and delivering weekly, members can continuously measure their progress in perspective of mutual expectations.
12. **Big data does enrich the model:**
    The methodology seeks for a combination of information available from both within the organization as from external sources. A combination between both provides the best answer to requirements. Since automation is part of the methodology, data from both sources can get processed near real-time and deliver knowledge to end-users accordingly.
13. **Try to destroy before you deploy:**
    Before delivering the results teams need to do the analysis whether this model adds more value and potential risk. If not, improvements are obviously needed. Part of automating everything is bringing in rules that inform the team when delivered work goes beyond set boundaries. “What if…?” questions set the knowledge up for current and future success.
14. **Retrospective (process) and maintenance:**
    As a continuous subject of discussion, the process gets discussed to provide room for improvement where needed. Since the model welcomes Big Data, it provides data to the model with a potential need to deployed models and so, maintenance wins in importance. Faster changing requirements have the same impact. Since Data and Ops are well connected at any time, teams work on maintenance constantly.

Roles:

- **Data-focused members and Ops-focused members:**
  Data-focused members have their expertise mainly on the data-part of the activities within the KD-methodology. The more Ops-focused members have a great understanding on what information is needed and how the business can benefit from the knowledge discovered. All

team members are working towards defined goals and roles are flexible. If help is needed at one part of the process, all members provide help where they can. This improves team-dynamics, understanding and in the end efficiency and effectiveness.

- **Architect and DataOps Mentor:**
  As both are not being a members of the team, both have a consulting role for the teams. The architect mainly helps to create a connection between not only the goals of operations but also the strategy, processes and IT governance of the organization. The DataOps Mentor is a role that can get consulted to have the teams walk through the process steps effectively. Since the teams are fully responsible for their results, consultation in both roles is a team responsibility as well. Their success depends on the results of all DataOps teams together.

Adjusted design steps

0. **New Business Request: A defined starting point**
   In the literature review, the paradigm of knowledge discovery as a continuous activity rather than a process with defined finishing step is provided. Building further on this paradigm, DataOps lacks a finishing point as well. Obviously, the loop starts with a "business problem" or question. In that phase, the team discusses and agrees on the budget.

1. **Problem understanding**
   A new request can be seen as a need, in the model presented as a problem based on the fact that the need is not easily fulfilled. By selecting stakeholders involved with that need, the opportunity is presented to align them in order to get one defined problem and (a set of) objective(s) and goal(s) for the KD-efforts.

   Team members have to get a clear view about the 'why' of KD-efforts. Obviously, answers can vary from 'we just need that figure' to 'this contributes enormous – today and in following years'. Seeing the objective(s) and goal(s) in the business perspective provides the opportunity to transfer data into information efficiently and effectively.

2. **Business understanding**
   In this phase, we review the connections of the defined problem within the organization. This includes the process of defining stakeholders and sources of the data. Additionally, it defines the analytics capability along three dimensions: the analytical environment of the organization, the quality of the data on an organizational level and how individuals let their decisions be guided by analytics.

   <u>Iteration (1):</u>
   By getting a better view about the business and relationship between the business and the KD-goals, a need to come to a sharper problem statement is a plausible scenario since in the end, all stakeholders have to be aligned about the definition of the the problem and how to work towards a solution. Continuous change on *how* is welcome at any time, all stakeholders need to get involved on *why* before starting the "Data-part" and stay aligned afterwards.

3. **Data understanding**
   For the proposed model, KD-workers need to get a good understanding about the playing field, the quality of the data, to come from data to information.

4. **Data selection and preparation**
   Having a clear overview of the available data, one can select, transform and clean the data.

   Iteration (2):
   During the selection and preparation, lessons learned can lead to a need for better understanding of the data and their relationship.

5. **Modeling**
   Having several techniques available for the same defined problem, KD-workers will have to try several ways to generate conversion routines from data to information. Although modeling, just one step of the of the "Data-part", sounds technical, "Ops"-professionals can contribute to their counterparts by periodically providing input on relevance.

6. **Testing**
   Obviously, the diversity of models bring different outcomes, not all of which are relevant. To optimize the accuracy of the model(s) selected for the next step, evaluation, a possibility to iterate is provided.

   Iteration (3):
   Insights lead to new insights, e.g. about the available data(set). Developing a better understanding about the data leads to better decisions. This iteration ensures an efficient way within the "Data"-part of the DataOps methodology.

7. **Evaluation**
   KD-efforts are only relevant when they support business objectives. When not having the match between information needs and delivered information, efforts are not effective. To prevent a forced deployment with depressing results, one could iterate and redo the "Data-part" once more (iteration 4) or redo the problem understanding and move forwards from there.

   Iteration (4):
   Lessons learned from the evaluation phase help KD-workers get a better understanding of the need of the business. Since the needed information will likely change, data understanding needs to get redone in order to loop once more through the "Data"-part.

   Iteration (5):
   First results can be delivered after having done one loop. Evaluation will lead to a conclusion about the risks and benefits of the model presented. Having this beta-state gives the team a possibility to present and deliver without a need to have a forced deployment when the model has not the required quality. This iteration is the only iteration that can be taken more than once.

8. **Deployment**
   As with all the phases in the model, this phase especially is a duo-owned phase. By deploying the model, the business will probably have many questions, both practical and technical. The team member with expertise on the data-side can both respond and fetch ideas for further improvement or future projects.

New data observations

The model can be enriched by using new data observations. These new data observations contain both new information collected over time or and new information sources. After a day, 24 hours of new data can potentially be put into the model. When the work is automated, new results are delivered with high efficiency. An example can be the relation between employees during said day and the revenue generated.

New information sources or tables can provide new data observations as well. If a data supplier provides new data that has not been implemented in the model yet, the model can be enriched by testing the added value of this. If this adds value, this will make the model even more robust.

Maintenance?

As adopted in the fundamentals, maintenance is part of the process as well. DataOps takes the position of KD rather as a process of continuous delivery than as a project with a defined end-phase. Automation of the process delivers continuous flows of information, but the process needs to be robust in order to deal with new data observations and software- and model updates.

Adjusted model